Identifying Market Structure: A Deep Network Representation Learning of Social Engagement

Yi Yang
Assistant Professor
Hong Kong University of Science and Technology
LSK Business Building, RM 4041
Clear Water Bay, Sai Kung, HK
+852-23586359
imyiyang@ust.hk

Kunpeng Zhang
Assistant Professor
University of Maryland, College Park
4316 Van Munching Hall
7699 Mowatt Ln
College Park, MD, USA 20742
301-405-0702
kpzhang@umd.edu

P.K. Kannan
Professor of Marketing
University of Maryland, College Park
3445 Van Munching Hall
7699 Mowatt Ln
College Park, MD, USA 20742
301-405-9187
pkannan@umd.edu

November 15, 2019
Revised September 4, 2020

Identifying Market Structure: A Deep Network Representation Learning of Social Engagement

Abstract

With rapid technological developments, product-market boundaries have become more dynamic. Consequently, competition for products and services is emerging outside the product-market boundaries traditionally based on SIC and NAICS classification codes. Identifying these fluid product-market boundaries is critical for firms not only to compete effectively within a market, but also to identify lurking threats and latent opportunities. Extant methods using surveys on consumer perceptions or purchase data will be unable to identify the impact that a brand from outside the boundary may have on brands within a product-market. Newly available big data on social media engagement presents such an opportunity. We propose a deep network representation learning framework to capture latent relationships among thousands of brands and across many categories, using millions of social media users' brand engagement data. We build a heterogeneous brand-user network and then compress the network into a lower dimensional space using a deep Autoencoder technique. We validate our technique using a novel link-prediction method and visualize the learned representations pictorially. We illustrate how our method can capture the dynamic changes of product market boundaries using two well-known events: the acquisition of Whole Foods by Amazon and the introduction of the Model 3 by Tesla.

Keywords: AI, Deep Representation Learning, Social Media, Competitive Market Structure, Big Data

Introduction

Firms compete in a market to satisfy the specific needs of consumers in the market. The market and the competing products comprise a "product-market" with the boundary defining the brands competing within that market. Identifying the product-market boundary and examining the strength of competition between brands within the product-market has long been important issues with strategic implications for next-generation product design, product positioning, new customer acquisition, and pricing and promotion decisions. While in many categories the product-market is stable over time, with managers (and regulators) usually focusing on brands within the boundaries from a competitive perspective, this is increasingly not the case. With the rapidly changing competitive environment ushered in by technological developments, the product-market boundaries themselves are changing and competitive threats and opportunities may emerge outside of the narrowly defined product-market boundaries. For example, the digital camera product-market has been upended by developments in smartphone categories. Similarly, Tesla who initially entered the product-market of high-end automobiles with a different fuel technology, has rolled out products for the lower-end market changing the competition in that product-market. Amazon, hitherto an online platform, has entered offline markets with the purchase of Whole Foods. In many such situations, product-market boundaries based on traditional SIC and NAICS industry classification codes may be unable to spot new threats and opportunities. Managers now need to have valuable insights on the fluid product-market boundaries which could help them spot potential competitors and complements, identify cross-promotion strategies, and develop firm-level strategies. This is precisely what our paper seeks to provide using large-scale (over a hundred million) social media user engagement data ("likes" and "comments") spanning several thousands of brands in different product/service categories.

Over the years, academics and practitioners have contributed significantly to developing various methods to define and identify market structure (see review by Shugan 2016). These include survey-based methods such as brand concept maps (BCM) (John et al. 2006) and ZMET (Zaltman and Coulter 1995),

methodologies based on observational purchase data (e.g., brand switching) (Kannan and Sanchez 1994), consideration sets (Ringel and Skiera 2016), and scanner-based purchase data (Erdem 1996; Novak 1993; Shugan 1987). Within the online context, researchers have used unstructured user click streams (Moe 2006), online search logs (Kim et al. 2011, Ringel and Skiera 2016), and customer reviews (Lee and Bradlow 2011). Some of these methods use data from the bottom of the purchase funnel, such as evaluation and purchase stage data, and thus assume the product-market boundaries are pre-specified. Even if some of these methods use data from the top of the funnel at the awareness or pre-evaluation stage, such as forum discussions (Netzer et al. 2012) and hashtags (Nam et al. 2017), these papers define a product-market boundary first and then examine the competition ***within*** the pre-specified product-market to make these methods implementable. Thus, many of the methods will not be able to capture the changes that occur to the product-market boundaries and/or the impact that a brand from outside the boundary may have on brands within a product-market.

Our methodology based on deep representation of user-brand relationships at the top of the purchase funnel overcomes the above limitation by creating a more inclusive representation of brands as well as users using heterogeneous networks. Many extant studied in market structure (e.g., Urban, Johnson and Hauser 1984; Grover and Srinivasan 1987, Kannan and Sanchez 1994; Erdem 1996;) as well as those studies using big data technologies (e.g., Lee and Bradlow 2011, Netzer et al 2012, Ringel and Skiera 2016, Culotta and Cutler 2016, Gabel, Guhl and Klapper 2019) view the competing/complementary brands as homogeneous networks. That is, they specify the relationship between any two brands in the product-market using metrics such as similarities or distances derived from brand switching, co-occurrences, word embeddings, etc. without directly modeling the entities (customers, individual consideration sets or individual reviews) that give rise to such similarities or distances. Our methodology based on heterogeneous networks considers both brands and users as primitives and uses as input the relationship in terms each user of liking and commenting on brands. This difference in homogeneous and heterogeneous networks imply that extant research considers aggregate data of relationships between

brands as input, while in our methodology we consider the disaggregate individual level relationships between users and brands as input.

The difference between a homogeneous network and a heterogeneous network also becomes more salient when a product-market boundary is not pre-specified. Consider, for example, a user who likes the brands United Airlines, Southwest and Hyatt Hotels. A homogeneous network would translate this data into three separate brand links between (1) United and Southwest, (2) Southwest and Hyatt, and (3) United and Hyatt. Our methodology based on heterogeneous networks allows implicit relationships between (1) users and brands, (2) brands and brands and (3) users and users to be leveraged in creating the multidimensional space and locate the three brands closer to this user. The additional disaggregate relationship information produces more accurate representations of the brands in the multidimensional space as compared to a homogeneous network. This additional information is discarded when the product-market boundary is pre-specified. Consider, for example, User 1 who likes United and Hyatt while User 2 likes Southwest and Hyatt. When the product-market is pre-specified as "airlines brands," information about the users liking the Hyatt brand is discarded. As a result, a piece of information that can provide insights into relationship between United and Southwest through their relationships with Hyatt is not considered. However, when we do not pre-specify the product-market boundaries as in our methodology we are able to leverage all such information in creating a very accurate representation of the brands.

The essence of our deep network representation learning based methodology is to represent brands and users in the same multidimensional "embedding" space. In other words, the more cross-industries engagement information we have from the users, the better we learn the user embeddings and, in turn, the better we learn the brand embeddings. Limiting product-market boundary to "airlines" would limit all engagement activities to within airline brands, and users engaging with brands from other industries are discarded. Our methodology avoids this problem and with the scaling power of our methodology we are

able to examine the relationships across 5000-plus brands based on the heterogeneous network of brands and users, which allows us to identify opportunities and threats for brands coming from outside of their traditional product-market boundaries. Even if a methodology based on homogeneous network is highly scalable as ours (e.g., Gabel, Guhl, and Klapper 2019) and thus allows boundaries to be not pre-specified, our methodology with heterogenous networks perform better as we show in our validation exercises.

Unlike the extant methods for identifying market structure that use data from consumers' lower funnel activities (purchase data, brand switching, price comparison data, consideration data) that pre-specify boundaries, we use upper funnel user-brand engagement data (such as liking and commenting on brand posts) from social media that spans product-markets. Our methodology used this user-brand engagement data to identify latent relations among a large number of brands. Previous research has documented that a user interacting with a brand online shows a brand affinity (Kuksov et al. 2013, Naylor et al. 2012), and can lead to (offline) purchase (Pereira et al. 2014), or express a desire to hurt the brand in favor of its rival brand (Ilhan et al. 2018). Therefore, we make the minimal assumption that if a user interacts with two brands, for example, Samsung and an HTC phone, it indicates the user has some level of interest -- greater than awareness -- in both brands. Simply by sharing interested users, the two brands are related to one another on a spectrum ranging from substitutes to independent to complements. If such patterns exist after observing activities on various brands from a large group of users (which could be millions of users on a social media platform), we argue that such a pair of brands have latent relations on some dimension, as learned from the user-brand data.

Based on the above premise, we first construct a large-scale heterogeneous user-brand network based on user engagement on brands' social media public fan pages. Then, we propose a deep network representation learning method to discover relationships within the data. Specifically, we use a deep learning method suitable for (1) handling large data efficiently and (2) learning complex patterns from data effectively (cf. Agrawal et al. 2018; LeCun et al. 2015; Timoshenko and Hauser 2019). The process

leads to a low-dimensional representation (i.e., a vector) for each brand and each user by training a deep Autoencoder on the network data. The deep Autoencoder is similar to traditional dimensionality reduction methods such as Principal Component Analysis in capturing latent factors in data with few dimensions. It is however very different from those methods in that it uses a non-linear transformation function to understand the latent patterns in data and at the same time reduce the noise in the data. In our context, the deep Autoencoder can preserve the first-order (user-brand direct connection) and the second-order (two users connecting to the same brand, or one user connecting to two different brands) network topology so that brands with network structural equivalence are located closer in the representation space, while brands with dissimilar network structures are located further away. This method also projects users and brands onto the same dimensional space, which can be used for many different follow-up analyses. For example, in this study we apply state-of-the-art visualization tools such as *t-SNE* (Maaten and Hinton 2008) to the learned brand vectors to visualize the product-markets characterizing the brands. Moreover, the proposed framework can capture changes in product-market boundaries by constructing a sequence of networks across different time frames to understand the dynamics of market structure.

We validate the product-markets identified through our methodology using a link prediction method, where using a calibration sample we learn the network representation and use it to predict the network structure in a validation sample. The results show that our proposed approach significantly outperforms several baselines on two standard metrics of predicting user-brand engagement on out-of-sample data. We also establish the face validity of the results through the identification of product-market boundaries. Our analysis of the user-brand engagement data of over five thousand brands and nearly 26 million users reveals product-market boundaries with high face validity – grouping of specific categories, high-end brands, and overlaps. Our event studies on Amazon's acquisition of Whole Foods and Tesla introducing the Model 3 illustrate how our methodology captures the changes in product-markets associated with these events. We also discuss how the market structure maps can reveal opportunities and threats facing a brand.

In summary, contributions of our paper include leveraging the information embedded in big data of user-brand engagement networks to identify product-markets without having to pre-specify boundaries. Using user-brand engagement heterogeneous network data at a much higher level in the purchase funnel (interest phase) and deep learning techniques provide us with insights at this scale and level of detail much better then extant methods. Our paper is among the first to apply deep network representation learning implemented using deep Autoencoder to social media data and show its usefulness for market structure discovery. Our ability to pin a large number of brands on the market structure map to precisely visualize brand relationships using the learned vector representations allows managers to identify opportunities and threats that lie beyond product-market boundaries. We provide illustrations of such insights and show product-market boundaries change as a function of events, which could be of use not only to managers within the product-markets but also to regulators trying to understand competition in markets. Finally, our study is apt illustration of how Artificial Intelligence (AI) can be used to better tackle a traditional marketing problem and provide insights. It is well known that three elements render AI techniques possible for-life applications: data, algorithm, and computing power (Agrawal et al. 2018). In this paper, we leverage deep learning and a network representation learning (algorithm) to understand market structure using a large-scale social media data (data). The model implementation is efficient under Nvidia P100 GPU, with Tensorflow as the backend framework (computing power).

Background and Theoretical Foundation

Extant work in identifying competitive market structures dates to the 1970s (e.g., Kalwani and Morrison 1977; Day, Shocker and Srivastava 1979) when diary-panel based brand-switching purchase data and survey-based consumer judgments of substitution-in-use or similarities were used to construct market structure maps. Developments since then have been based on the availability of the volume and variety of data and methodology in terms of their sophistication and capability to handle large volumes of data. We briefly review them from the two perspectives of data and methodology.

Data

Early studies depended on customer data generated either at a late stage of the customer journey or the very beginning of the journey. For example, purchase data collected using diary-panels or survey of brand perceptions – judgements of similarities or substitution-in-use – were commonly used for constructing brand-switching data or perceptual maps of brand relationships. The increased availability of scanner-panel data of purchases, market structure models with marketing mix (e.g., Carpenter and Lehmann 1985; Kannan and Wright 1991), and dynamic market structure models (e.g., Erdem 1996) provided more detailed insights into inter-brand relationships and competition. Focusing on the early stages of the customer journey, approaches such brand concept maps (BCM) (John et al. 2006) and ZMET (Zaltman and Coulter 1995) relied on data collected using surveys and, therefore, were effort intensive. Given the scaling issues with the MLE-based models and the limitations with survey data, the market definition problem was ignored, and product-market boundaries were pre-specified generally at the industry level so that a smaller number of brands within an industry could be analyzed.

The advent of online sources, such as review platforms, social media platforms, and clickstream data, has dramatically increased the volume and variety of data for market structure studies, especially at the awareness, search, and consideration stages of the customer journey. For example, the study by Kim et al. (2011) relies on Amazon's customer search logs on camcorders to derive market structure. Lee and Bradlow (2011) visualize competitive market structure in the digital camera industry using user-generated online customer reviews that mainly comment on product attributes and brands' relative positions. The study by Netzer et al. (2012) relies on data from online discussion forums to build a market structure of the automobile industry using a hybrid text mining and network analysis method. Similarly, Ringel and Skiera (2016) use search history from a product and price comparison site to derive customers' consideration sets that reflect competition among LED-TVs at the SKU level. It is important to note that

even with a large volume of data, these studies pre-define the product-market boundaries at the industry level to make the analyses viable.

There are other studies where the product-market boundaries are not pre-defined. For example, France and Ghose (2016) introduce a method for identifying, analyzing, and visualizing sub-markets in product categories from online reviews. Nam, Joshi, and Kannan (2017) use hashtags from a social tagging website to infer brand relations across categories. Similarly, Culotta and Cutler (2016) propose to extract brand-related attributes and build brand conception maps using hashtags from Twitter, where such pre-defined boundaries are not necessary. More recently, Gabel, Guhl and Klapper (2019) analyze customers' market baskets of items purchased on shopping trips using word embeddings. (However, from a methodology perspective all these studies use homogeneous networks – a distinct disadvantage as we have discussed earlier). Our proposed methodology focuses on the early stages of the customer journey and allows the product-market boundaries to emerge from the data. More importantly, the scale at which we analyze that data, which is much larger than any of the extant methods (except for Gabel, Guhl and Klapper 2019), is key to analyzing the relationships spanning multiple categories. Our proposed method is capable of handling thousands of brands and millions of users. Table 1 summarizes the studies based on the type of data used.

<Insert Table 1 Here>

Methodologies

Much of the online data generated in the early stages of the customer journey tend to be unstructured (online reviews, social-tags, hashtags, etc.). In order to extract product names or attributes included in data, researchers develop and apply various text mining-based technologies to reviews, discussions, and summaries in online forums. For example, Netzer et al. (2012) propose a model combining conditional random fields (CRF) and predefined linguistic rules to extract product keywords. Pant and Sheng (2015) focus on firm-generated content in websites, compute firm similarity based on textual descriptions (from

the first few pages of firm websites), and use TF-IDF (term frequency-inverse document frequency) and website structures link analysis (in-links and out-links) to derive firm competing relationship. Ringel and Skiera (2016) construct consideration sets from consumer co-search data, while Gabel, Guhl and Klapper (2019) use word embeddings. Our method is significantly different from these methodologies as the above methodologies are based on homogeneous networks while we use heterogeneous networks. Table 2 summarizes selected extent studies on market structure analysis in multiple dimensions and highlights the positives of our proposed methodology.

<Insert Table 2 Here>

Social Media Engagement

Our proposed methodology analyzes social media engagement data in the form of user-brand links. Many social media platforms such as Facebook, Twitter, and Instagram host public fan pages created by firms to facilitate the communication with customers and promote products. The user-brand engagement could be in the form of a user "liking" a post by the brand, "sharing" a brand post, or "commenting" on a brand post. Since each of these "likes," "shares," and "comments/posts" is a user-brand link in our study, it is important to understand what they represent. Surveys of fans of brands have revealed many reasons as to why users "like" a brand or post/share comments. These include positive attitude such as: to support a brand they like, to get a coupon or discount, to get regular updates from the brands they like, to participate in contests, to share personal experiences, to share their interests/lifestyles with others, to research brands, to imitate a friend who likes the brand, to act on a recommendation from another fan, etc. (Kuksov et al. 2013, Naylor et al. 2012, Pelletier and Horky 2015, Pereira et al. 2014). In contrast to the positive sentiments, users may also leave negative comments to hurt brand in favor of its rival brand (Ilhan et al. 2018). In our proposed approach, we make a minimal assumption by creating a user-brand link regardless of the type of engagement. This minimal assumption is based upon a rationale that users interacting with a brand online exhibit their interest towards the brand to some extent. Thus, the two brands are related to

one another on a spectrum ranging from substitutes to independent to complements. Prior research has examined such contexts and studied the impact of user engagement on brand image and customer purchase intentions with mixed results (De Vries, Gensler, and Leeflang 2012; Lipsman et al. 2012; Naylor, Lamberton, and West 2012; Goh, Heng, and Lin 2013; Hoffman, Novak, and Kang 2017). For example, Goh, Heng, and Lin (2013) find that user engagement in social media brand communities leads to a positive increase in purchase expenditures. Mochon et al. (2017) use a field experiment to find that users who liked a gym brand online were likely to become members of that gym offline. It is not the organic liking by the user that lead to the offline purchase; more often "liking" a Facebook page is used as a platform for firm-initiated promotional communications. In another field experiment setting, John et al. (2017) find that "liking" is simply a symptom of a positive brand attitude and does not imply the fan is any more loyal to the brand or any more likely to purchase the brand. Additionally, it is only when users who liked the brand are targeted using promotional communication by the firm that purchase probabilities increase. Thus, for our research purposes we will treat a "like," "share," or a "comment/post" as exhibiting an interest towards the brand at the beginning of the customer journey. Such a tendency for users to connect to brands is generally interpreted as interest, and reflects possibly broader (e.g., offline) interactions (Culotta and Cutler 2016, Kuksov et al. 2013, Naylor et al. 2012, Netzer et al. 2012), which is consistent with our treatment. Our proposed approach is also consistent with research in social network analysis suggesting that social network structure equivalence reflects value/interest homophily and can be used to measure social proximity (McPherson, Smith-Lovin, and Cook 2001).

Methodology

The social network platforms, such as Facebook, Instagram, and Twitter, can be abstracted as a network containing business (firm) accounts and individual user accounts. The public fan pages of business accounts are used by firms to communicate with their customers and fans. Users interact with brands and with each other in different ways, such as commenting, liking, sharing, and following. To discover latent

relationships among brands, we propose a deep network representation learning framework which has the following steps as summarized as in Figure 1.

<Insert Figure 1 Here>

Step 1: Data Collection. We specify a set of brands that is of interest in the social network platform. We then download all available user engagement data from the brands' public fan pages during the appropriate time window depending on managerial interest. A user engagement is defined as either liking or commenting on a firm's post on its public fan page. Note that for the sake of privacy, we do not attempt to collect any personal information of users. Rather, the only thing we obtain is the unique user identifier, assigned by Facebook, and their public engagement activities, which is consistent with recent study on social media marketing (Ilhan et al. 2018, Kübler et al. 2019)[1]. Moreover, different platforms may have their own specific data policy. For example, collecting individuals who liked a given page is not permitted by Facebook. Such data restrictions and potential ethical concerns do come at a research cost as we would not be able to verify how representative they are of the population at large. Therefore, developing a sophisticated model becomes necessary to analyze publicly available data.

Step 2: Network Construction. We start with a cleansing operation to remove spurious users. We then construct a heterogeneous user-brand network including all selected brands and all users engaging with them. A brand node and a user node are connected if the user engages with the brand. The strength of an edge between a brand node and a user node is the engagement frequency.

Step 3: Deep Network Representation Learning. The deep network representation learning algorithm represents each node (brand or user) as a low-dimensional vector, also known as a node embedding. Embedding techniques are not new in marketing. For example, Timoshenko and Hauser (2019) adopt pre-trained word embeddings, where each word is represented as a low-dimensional vector,

[1] In fact, obtaining detailed user personal information for marketing analysis (e.g., political targeting) is controversial and subject to ethical concerns, such as the Facebook–Cambridge Analytica data scandal.

to extract insights from textual reviews. However, our node embeddings are trained via an unsupervised deep Autoencoder. This representation learning is essential to data-driven analysis, and the learned low-dimensional embeddings are useful for the downstream task of identifying and visualizing the product-markets.

The objective in using an Autoencoder is to learn the representation of the data so that each node can be represented in a lower dimensional space while the network structure between users and brands is preserved. It trains the network to ignore the "noise" in the data and focus on the primary latent structure. The Autoencoder reduces the dimensionality of the input data to a "bottleneck" (the reduced encoding), and using the reduced encoding as input, reconstructs a representation of the original data. Learning occurs through backpropagation of the loss (see detailed definition in Appendix) to get the reconstructed representation as close as possible to the original representation while eliminating noise. It is the bottleneck reduced encoding we are interested in for developing market structure. In essence, we can compare the dimensionality reduction functionality of the Autoencoder with that of Principal Component Analysis (PCA). While in PCA the reduced dimensions are linear combinations of the input variables, the reduced dimensions in Autoencoder are non-linear and non-orthogonal achieved through non-linear activations of the neurons allowing the model to learn more powerful generalizations than what PCA can.

In our application, the Autoencoder works on the large heterogeneous network in an attempt to preserve the network structure such that (i) nodes directly connected have similar vectors (closer to each other) in the reduced embedding space, and (ii) nodes that are not directly connected but share structural equivalence (such as many common neighbors) are also similar in the embedding space. These two types of similarity are referred to as the first-order (direct connection) similarity and the second-order (network structural equivalence) similarity. Formally, we denote an aforementioned network as $G = (V^b, V^u, E)$, where $V^b = (v_1^b, v_2^b, \dots, v_n^b)$ represents a set of n brand nodes, $V^u = (v_1^u, v_2^u, \dots, v_m^u)$ represents m user nodes, and $E = \{e_{i,j}\}, i \leq m, j \leq n$ represents all links between users and brands. $e_{i,j}$ indicates an

engagement between user i and brand j. Given such a network G, the network representation aims to learn a mapping function $f: v_i^b, v_j^u \mapsto w_i^b, w_j^u \in R^d$, where $d \ll \min(m, n)$. w_i^b, w_j^u are called brand embedding and user embedding, respectively. A commonly used embedding dimensionality d is 300 (Mikolov, Chen, et al. 2013, Mikolov, Sutskever, et al. 2013). The objective of the mapping function is to develop appropriate embeddings so that the brand proximities, brand-user proximities, and user proximities exhibited in the original network are preserved as much as possible in the reduced embedding space. Technical details of the Autoencoder methodology, and parameter tuning are discussed in Appendix A1.

Prior research of network analysis relies on network adjacent matrix representation, that is, a brand node is represented as a $|V^u|$-dimension vector where $|V^u|$ is the number of unique user nodes in the network. Each element in the vector corresponds to a user. If the user at a particular index has a connection with the brand in the network, that element is marked as 1 and 0 otherwise. The brand vector is usually extremely sparse given the fact that each brand only engages with a small subset of users. Similarly, the user vector is also very sparse since each user only engages with a small subset of brands. Using this representation to measure similarity is inaccurate – not mention inefficient – for such an order of magnitude. In contrast, representing brands as dense low-dimensional vectors allows us to capture brand relations from multiple facets, and we use a toy example (shown in Figure 2) to illustrate how network representation learning works.

<Insert Figure 2 Here>

Suppose we have three brands (B1, B2, B3) and five users (U1, U2, U3, U4, U5) in a network. Representation learning aims to find a mapping function so that each node is represented as a low-dimensional vector (for the sake of this illustration let us assume that it is 3-dimensional). The mapping function is optimal when nodes exhibiting similar structures (first-order and/(or) second order) are projected onto similar vectors in the reduced embedding space (assumed to be 3-dimensional space in this

illustration). Since U1 engaged with B1, we expect the vector representation of B1 and U1 to be close. Similarly, B2's representation is closer to B1's representation than to B3's as B2 shares more users with B1 than with B3. Since B2 has some additional network structure, such as connections with U4 and U5, its representation leans towards U4 and U5. All representations are jointly learned in a similar way.

Step 4: Market Structure Discovery. Once we obtain vector representation for brands and users, we can use learned embeddings to efficiently compute similarity among brands and to visualize natural clusters of related brands. Finding similar brands to a focal brand can be achieved by a nearest neighbor search based on the widely used cosine similarity. Cosine similarity measures the cosine of the angle between two vectors and has a range [-1, 1][2]. Visualizing natural clusters of related brands can be achieved by a dimension reduction method, such as *t-SNE* (Maaten and Hinton 2008), which projects high-dimensional data into a low-dimensional space (e.g., two or three dimensions). It has been used for visualization in a wide range of applications and is especially well-suited for visualizing high-dimensional representations learned from deep neural networks. *t-SNE* preserves the distance of data points well such that data points nearby in a high-dimensional space would be close in a lower dimensional space, while distant data points would be further apart in a lower dimensional space. Specifically, the input of *t-SNE* is the learned vectors from our network representation learning in the reduced dimension space with $d=300$ and the output is the vectors with 2 dimensions. Thus, we shall observe that related brands are surrounding each other in the reduced 2-dimensional space after *t-SNE*. Note that other high-dimensional visualization method such as UMAP (McInnes et al. 2018) can also be applied to visualize the derived market structure. Both *t-SNE* and UMAP are designed to provide a very informative visualization of heterogeneity in data. They have comparable quality of preserving global and local structure.

[2] Note that we do not use Jaccard similarity because the brand vector space is a continuous space and Jaccard similarity is specifically designed for a discrete space, nor Euclidian distance due to its poor performance in a high dimensional space.

Data

In this study, we use Facebook as our empirical benchmark, as it is one of the largest and most representative online social network platforms. Note that our model can be generalized to other similar social network platforms.

To collect Facebook data, we first obtain a list of U.S. brands with the most followers from the social media marketing website Socialbakers[3]. Facebook public fan pages are categorized into several groups on Socialbakers, such as Brands, Celebrities, Community, Entertainment, Media, Place, Society, and Sport. Without loss of generality, we focus on the "Brands" as it covers a wide range of different industries and is more interesting to marketers. On Facebook, every brand is associated with a category that is chosen from the predefined Facebook system when creating the public page. This category label is solely determined by the brand and is aligned with its core business. For example, Walmart is under the category of "*retail*," and Amazon.com is under the "*ecommerce*" category. In total, we obtain 5,478 different brands, covering 25 different categories. The largest brand, in terms of number of followers, is Walmart, with 30 million followers. The smallest brand, is Bladz Jewelry in the "*fashion*" category, with 100 thousand followers. Figure 3 shows the histogram of number of followers of brand Facebook page. We observe that the dataset contains a variety of brands with varying popularity, which makes us believe that this dataset is representative of brands on Facebook.

<Insert Figure 3 Here>

On Facebook, firms post on their public fan pages and allow users to comment, like, and share. The posts become an important marketing channel for businesses to interact with their customers. We use Facebook Graph API[4] to download all activities visible on a brand page such as posts by the brand

[3] Socialbakers is a global AI-powered social media marketing company offering a marketing software-as-a-service platform called the Socialbakers Suite. It includes data from Facebook, Twitter, and YouTube. https://www.socialbakers.com/.

[4] https://developers.facebook.com/docs/graph-api/

administrator, as well as posts by users, including comments and likes on brand posts. Facebook added more reaction emotions such as 'love,' 'haha,' 'wow,' 'sad,' and 'angry' in 2016. Our dataset does not include these reaction emotions because the Graph API returns only the 'likes.' Moreover, users can 'share' brands' posts, but the Graph API does not provide individual level data regarding who shares which post, rather it provides an overall share count for each post. Therefore, our dataset does not include 'share' engagement. It is worth emphasizing that to ensure privacy protection, we do not download any user profile information nor examine the content of user comments. All engagement activities are represented by unique user identifiers, regardless of whether the user has a public or private Facebook profile, and brand identifiers. The dataset collected for this study covers the duration from January 1, 2017, through January 1, 2018. In total, we obtain 106,580,172 user-brand engagement activities from 25,992,832 unique users. Since prior research has shown that online interaction is a reflection of broader and even offline interaction (Pauwels and Van Ewijk 2013), given the scale of user online engagement in this study, we believe it is a good proxy of how the overall consumer population perceives these analyzed brands.

Data cleaning. To ensure data quality and robust results, we design a set of rules to remove fake users and their corresponding activities. As fake accounts and fraudulent activities have become more pervasive, researchers and social media firms are increasingly paying more attention to these problems (Mukherjee et al. 2012, Van Vlasselaer et al. 2016, Zahedi et al. 2015). For comments on Facebook brand pages to reflect genuine user experiences, opinions, and interactions with brands, such fraudulent activities should be detected and removed. Following prior work on Facebook (Zhang et al. 2016), we replicate a set of similar rules to remove fake users and their posts. For example, we do find one user who liked posts across 475 different brands. As most users are likely to be interested in few brands, we remove users who like posts on more than 200 brands, which accounts for 0.01% of the total users, and 1.6% of the total user-brand engagement. We also remove users who posted duplicate comments containing URL links. Table 3 describes the resulting data using a heterogeneous user-brand network. The brands' degree

distribution (number of connections) exhibits a scale-free distribution (shown in Figure 4), a well-documented phenomenon in most social networks.

<Insert Table 3 Here>

<Insert Figure 4 Here>

Evaluation and Results

In this section, we first introduce our design to quantitatively evaluate our proposed methodology and compare its performance with several baselines. Then we present the market structure derived from our learned brand representation, visualized using a sophisticated dimension reduction technique *t-SNE*.

Evaluation using Link Prediction

In studying market structure, there is lack of ground truth about the identified structure, that is, knowledge of what the "true" structure is. As a result, demonstrating the performance of various proposed methods is challenging. One may argue that the industry classification (e.g., SIC or NAICS) of brands can be used for evaluation and face validity for the results. However, these classification systems are static, do not re-classify firms as the product-market evolves, and, therefore, are unable to accommodate innovations that create entirely new product markets (Bhojraj et al. 2003, Hoberg and Phillips 2016, Jacobs and O'Neill 2003). To address the challenge, we propose an alternative and novel way to evaluate the identified market structure. An identified market structure is a function of the brand representation and so an accurate representation is more likely to identify valid market structures. This approach is supported by prior research showing a strong relationship between brand image and the characteristics of brand's supporters and followers (Naylor et al. 2012, Kuksov et al. 2013, Culotta and Cutler 2016). If a network learning method is capable of accurately representing network nodes accounting for these relationships between brands and users, then it would be able to predict the future links between brands and users accurately. Therefore, we use a cross-validation procedure under a *link prediction research design*, where we predict the most likely formed links of user-brand engagement in an out-of-sample network given the

brand vectors and user vectors learned from a training network. We use the user-brand interactions from the first half of time span in our data to build a training network ($G_{0,1}$) and use the second half to build a testing network ($G_{1,2}$). $G_{1,2}$ has 7,247,410 links (1,996,354 new links), formed by 1,547,762 users and 1,511 brands (also in the training network). The likelihood of a link formation is measured by the proximity of a learned brand vector and a learned user vector. Note that link prediction performance is significantly correlated with the quality of learned vectors, given the assumption that a better network representation learning can predict new interactions between users and brands with a high accuracy.

To demonstrate the superiority of our proposed method, we compile a set of representative baselines. Specifically, we compile a 2x2 research design with two different network structures (homogeneous vs. heterogeneous) and two different algorithms (shallow model vs. deep model). We follow prior literature in social network link prediction (Liben-Nowell and Kleinberg 2007) and use precision-recall as evaluation metrics (see details in Appendix A2). Note that the accuracy of a link prediction under a random strategy is approximately 0.085%.

Overall, our analysis shows that (i) link prediction using representation learned from our heterogeneous user-brand network performs better than a reduced homogenous network – a widely used method by extant approaches; (ii) deep learning-based methods learn better representation than shallow machine learning methods; and (iii) our deep learning-based model is robust and able to handle sparse networks as compared to baselines. Table 4 and Table 5 summarize the performance for the case where we randomly select 100 and 1,000 users, respectively. Note that all *p*-values in parentheses are obtained under a *t*-test from 10 runs of every model. We can see that our method significantly outperforms baselines in both *precision@k* and *recall@k* at all different *k*'s. As an illustration, consider Column 4 (with k = 1,000) in Table 4, the shallow model on the homogeneous brand-brand network has precision of 0.078 which means that only 7.8% of the predicted links are actually formed during the testing period. In contrast, our deep model on the same network brings a slight improvement to 0.082. This suggests that

shallow and deep models have comparable performance when data is small and homogeneous. We then apply our deep model to the heterogeneous network, which significantly improves the precision by approximately 58.9% over traditional methods (0.124 vs. 0.078). We observe similar trends for metric *recall@k*. For *k*=1,000, shallow and deep models on the homogeneous network are able to retrieve 60.2% and 68.6% links, respectively (Table 4, Column 4). Further, *precision@k* decreases as *k* becomes larger, while *recall@k* increases. When *k* is large, many false links are predicted as well as true links. This reflects the famous precision-recall tradeoff that any model can be adjusted to improve precision at the expense of recall, or vice versa.

By further investigating Table 4 (*N*=100) with Table 5 (*N*=1,000), we find that the *precision* is higher and the *recall* is lower when the number of selected users is large. This is expected because we have higher chances to select true positive links when the number of users increases. On the other hand, the total number of true links in the testing network also increases by a magnitude, and thus the *recall* decreases.

<Insert Table 4 Here>

<Insert Table 5 Here>

To study the impact of training size on performance, we vary the training size with different network sparsity. We randomly remove a certain percentage of links from the training network and learn representation of users and brands. Then we predict user-brand links and measure the *precision* and *recall* using the out-of-sample testing network. As we can see in Table 6, our method still significantly outperforms baselines, especially when network sparsity is extremely high. For example, our method improves the *precision*@1000 by 77.7% (0.183 vs. 0.103) and *recall*@1000 by 55.0% (0.080 vs. 0.124), when only 10% links are kept in the training network. This suggests that our method handles sparsity better than baselines, which is very important since most real-world networks are extremely sparse.

<Insert Table 6 Here>

Visualization of Market Structure

With the learned brand representation vectors, we can visualize how the brands are grouped from a global perspective, and even zoom-in to examine local fine-grained brand relationships. We use *t-SNE* to obtain market structure visualization. *t-SNE* (Maaten and Hinton 2008), a dimension reduction technique, has been shown to preserve global structure better than the multi-dimensional scaling (MDS) (e.g. Kim, Albuquerque, and Bronnenberg 2011). In our context, we use *t-SNE* on the learned 300-dimensional brand representations to obtain the associated 2-dimensional visualization map[5]. Figure 5 presents the global structure of the brands in our Facebook data. Each data point in the figure denotes a brand belonging to one of the 25 categories, and each category is indicated by a different color. We can interpret the visualization as follows: the closer any two brands are in the figure, the more similar their brand representations are in the 300-dimensional space (see Figure 5). The color codes in the map indicate brands in the same Facebook category, with the category label self-identified by the brands themselves on Facebook.

<Insert Figure 5 Here>

There are several observations from the global Facebook brand market structure map. First, there are clear grouping patterns into clusters, particularly between brands in the same industry (points with same color tend to be in a group). For example, Cluster 1 in Figure 5 includes non-luxury domestic and imported automobile brands such as Toyota, Nissan, Mazda, as well as some automobile accessories brands such as Michelin, DENSO, and Auto Parts. Note that in our data we have several luxury car brands such as BMW, Mercedes-Benz, Audi, Tesla, and Maserati, which are not close to the brands in Cluster 1. In fact, they are clustered in a different region of the map with other luxury brands such as

[5] *t-SNE* aims to minimize the Kullback-Leibler divergence (KL-Divergence) between the probability distribution over pairs of data points in the original high-dimensional space and that in the reduced dimension space. It involves some hyperparameters, which are tuned based on the model perplexity measured by the Bayesian Information Criteria (BIC) (Schwarz 1978). In our analysis, we choose a variant of BIC (Cao and Wang 2017). We find the perplexity score of 40 in a range of 10 – 50 achieves the best BIC criteria. Other hyperparameters, such as learning rate and the number of iterations are set to the values when the KL-divergence gets converged.

Channel, Gucci, Cartier, and others. Such a separation between luxury car brands and non-luxury car brands further confirms that brand representation learned from our approach captures latent semantics in multiple dimensions, not only on the industry dimension but also on the price dimension. Second, some brands appear in categories which are different than what one would normally expect. There are two explanations for this mis-categorization (in addition to the one related to the latent semantics captured in the luxury-non-luxury automobile case). First, our category label is obtained from the self-identified category label indicated by each brand on Facebook. For example, Amazon lists itself in the e-commerce category instead of the e-commerce category while Apple's chosen category is service rather than electronics. Second, each brand might have several businesses across different categories. For example, Amazon's business is related to the high-tech, shipping and delivery industry, as well as retail and supermarket (after acquiring Whole Foods) industries. The strength of our methodology lies in capturing these relationships into a single map given the ease with which it locates thousands of brands in the market structure map, thereby highlighting the complex and possibly overlapping product-market boundaries characterizing these brands.

To further examine the specifics of the product-market boundaries, we zoom-in on the four clusters in the figure to examine the fine-grained local market structures, which are displayed in Figure 6. Subfigure 1 which we already discussed displays automobile brands along with automobile accessories and motorcycle brands at the top. Subfigure 2 displays premium vacation resort brands, such as The Signature at MGM Grand and the Coconut Bay Beach Resort & Spa. Subfigure 3 and Subfigure 4 contain airline brands and cosmetic brands, respectively. Taken together these maps provide face validity to our methodology in terms of core brands making up an industry and the overlaps among product-markets.

<Insert Figure 6 Here>

Identifying Proximal Brands

While visual mapping is sufficient to provide a gestalt picture of all the five thousand plus brands in the aggregate, it does not provide the actual distance between the brand vectors in the reduced dimension space. Since identifying proximal brands for substitute/complement analysis is a critical task in marketing decisions (Day et al. 1979), we focus on identifying proximal brands from the perspective of a focal brand. In doing so, we offer a new perspective that reflects the nature of relationship ranging from substitutes to complements in the social network space.

In this illustration, we choose United Airlines and Southwest Airlines from the airlines category and Audi USA and Nissan from the automobile category, as these brands are generally regarded as having different consumer bases and belonging to different sub-markets. Each of the four brands is referred to as a focal brand, and we find their top-10 proximal brands based on cosine similarity. From the results in Table 7 we can obtain several interesting insights. First, our method is able to capture specific brand latent characteristics. For example, Southwest Airlines is generally considered as a low-budget airline compared to United. The brands most proximal to Southwest Airlines and United reflect this difference. The proximal brands for Southwest Airlines are JetBlue, Frontier Airline, and Allegiant, while the most proximal brands for United are major domestic and international airlines, such as American Airlines, Delta, Lufthansa, All Nippon, Air China, LATAM Airlines, and Air New Zealand. Similar results also are identified in the automobile industry. Top proximal brands to Audi USA are Mercedes-Benz USA and BMW USA, which are generally high-end luxury car brands. In contrast, Nissan is closer to Mazda, Toyota, and Volkswagen, all of which produce more affordable cars.

Second, we also observe the asymmetric competition (cf., Ringel and Skiera 2016). Given a brand A and its top proximal brand set S_A, we can identify a set of brands S_B that is proximal to a brand B that appears in S_A. S_B need not necessarily contain A. And even if $A \in S_B$, the order of A in S_B might be very

different from the order of B in S_A. For example, Southwest Airlines is the fourth most proximal brand to United while United ranks sixth in the set of top proximal brands to Southwest Airlines.

Third, unlike prior market structure analysis where proximal brands are usually from the same industry as the focal brand, the top most proximal brands derived from our analysis are from different industries. For example, a brand called "Airfarewatchdog" is proximal to both United and Southwest Airlines. Airfarewatchdog is a deal-finder for flight tickets and has a large follower base (over 1 million) on Facebook. Traditional market analysis may simply ignore this brand, as it is not an airline. Further, it is also interesting to see that Southwest Airlines is closer to Airfarewatchdog than to United which may indicate that the fans of Southwest Airlines are more likely to use a deal finder before purchasing flight tickets; thus, Airfarewatchdog could be a complement to Southwest when customers look for cheap flights at that site and end up at Southwest, or it could potentially compete with Southwest. In either case, Southwest could focus more on this site and examine the nature of the relationship. Similarly, we see Kawasaki USA, an innovative motorsport vehicle manufacturer, is proximal to Audi USA. This cross-industry brand proximity very well demonstrates that representation learning can capture latent characteristics of brands and explore brand relationship from different perspectives. We believe this advantage can provide new insights to marketing analysis as we show next.

<Insert Table 7 Here>

Identifying Opportunities/Threats

Our market structure map can help managers identify brands outside of the product-market that are close to a specific brand within the product-market and thus identify opportunities and threats to different brands. Let's take the airlines (Subfigure 3) product-market as an example. Based on our analysis, Disney Cruise Line and Hyatt are two brands outside of the airlines product-market, but are identified as proximal brands to Southwest but not for United. These proximal locations simply are due to a greater number of users in our dataset liking both Southwest and Hyatt (2709) versus number of those users

liking both United and Hyatt (954). Similarly, a greater number of users like both Southwest and Disney Cruise (3050) than those liking both United and Disney Cruise (729). Our link prediction validation exercise where we divide the dataset into first half as the training set and the second-half as the hold-out for link prediction, also confirms the robustness of these proximal locations. In the hold-out set, there are 1,738 Disney Cruise Line users who also engage with Southwest, out of which 1,257 engaged with other brands, but not Southwest, in the training set. Since our methodology learns that Disney Cruise Line and Southwest Airlines have similar representations, as they are shown to be neighbors in Figure 5.3, it predicts a link between those 1,257 Disney Cruise Line users and Southwest than any randomly selected users, confirming that these proximal relationships we find are robust.

Such findings can provide opportunities for Southwest as they can target those who like Disney Cruise and Hyatt in social media. They can cross-promote these brands by teaming up with Disney Cruise and/or Hyatt on each other's websites. They can also launch coalition loyalty programs. From the viewpoint of other hotel chains who are competitors to Hyatt, these could be potential threats so getting such insights early on may help them take proactive actions. Such opportunities/threats are difficult to identify when product-markets are pre-specified, and they cannot be obtained easily through other means.

Large brand versus Small brand

Our user engagement dataset contains top 5,478 brands, ranked by their popularity (number of followers as of data collection period) on Facebook which are primarily large brands. A key question is whether our proposed approach is still able to identify meaningful market structure for smaller brands. Small brands have the potential to increase consumer awareness and interest towards them (Hanssens et al. 2014) if they can find right positions in the product-market structure, which could lead to permanent benefit in terms of them gaining a competitive advantage (Slotegraaf and Pauwels 2008). Therefore, to test whether our methodology is able to capture relationships among large brands as well as small and local business brands, we add a set of smaller brands to the original dataset. Specifically, we focus on the "Travel"

category as it includes many small local travel agencies, and their followers on Facebook range from a few hundreds to a few thousands on average. In total, we have 242 additional brands. Figure 7 plots the distribution of brand size of these additional brands. It shows that the travel brands are much smaller than those in the original dataset.

<Insert Figure 7 Here>

Upon applying our methodology to the enlarged dataset, we can observe (Figure 7) that these 242 travel brands are predominantly located in two areas. This pattern indicates that the latent brand relationship is well captured, even when brands have few engagement activities due to their smaller user bases. This result also highlights the advantage of our deep network representation learning method. This is a distinct advantage of our user-brand heterogenous network-based approach as brand representation can still be achieved via direct learning the first- and the second-order connectivity using deep learning. In a homogenous network such small number of shared user base could result in a failure to capture such proximal locations.

The market structure uncovered for these small businesses by identifying their proximal brands has good face validity. For example, "The Luxury Travel Expert" is an information portal for luxury travel and premium tours, with about 11,000 followers on Facebook, as of our data collection period. Most posts receive less than 10 comments and likes. The top 10 proximal brands based on the cosine similarity are: "*Smithsonian Journeys*," "*The Peninsula Beverly Hills*," "*Peter Sommer Travels*," "*Quasar Expeditions*," "*DuVine Cycling*," "*International Expeditions*," "*TCS World Travel*," "*Zegrahm Expeditions*," "*Liberty Helicopters*," "*Frosch Travel*." It is noteworthy to observe that they are also small travel brands, with focuses on expert-led, small-group, luxury and premium tours. The results further confirm that our deep network representation learning method is generalizable to both small and large brands.

Within-Industry Market Structure Analysis

Extant methods typically pre-define the product-market boundary to derive market structure and brand relationships. In contrast, we allow the product-market boundaries to emerge from data. Therefore, a natural question is whether it is necessary to have a broader range of brands from other industries to derive a high-quality market structure of a specific industry. While managers would typically focus on engagement data for their brands and for brands within the same industry, how does engagement data from brands in different categories help? To answer this question, we choose the "auto" category and only use the engagement data from the "auto" brands to derive the market structure. In the dataset, we have 163 "auto" brands, including cars and car accessories brands (such as tires, oil), and 2.7 million user engagements in total. The analysis shows (Figure 8) that structures with reasonable face validity still emerge using only the "auto" brands data. For example, the top left corner in Figure 8 (right) presents a cluster of imported auto brands such as Kia Motor America, Toyota, Nissan. However, compared against the derived "auto" brand market structure learned from using all brands data, as shown in Figure 6 (1), the market structure is less clustered and more ambiguous.

We now compare the market structure using the engagement data from the "auto" brands alone with that from all brands across categories in a qualitative manner. Specifically, we choose the brand "FMF Racing," which is a company that develops dirt bike exhausts for off-road or racing motocross riding. Using the engagement data from the "auto" brands alone, the top 10 proximal brands are "Lucas Oil," "KTM USA," "Yamaha Motor," "Arctic Cat," "Two Brothers Racing," "Phoenix Pro Scooters," "Auto Alliance," "Valvoline USA," "Lance Camper," "Castrol." Some are related to off-road motocross riding, while others are not. For example, "Lucas Oil," "Valvoline USA" and "Castrol" are global brands of automotive oil. "Auto Alliance" is a trade group of automobile manufacturers, and "Lance Camper" is a manufacture of travel trailers and truck campers.

In contrast, the top 10 proximal brands to "FMF Racing" emerging from using all categories data are "KTM USA," "Polaris Snowmobiles," "Fox Racing," "Mickey Thompson Performance Tires & Wheels," "Two Brothers Racing," "King Shocks," "Arctic Cat," "Addictive Desert Designs," "NISMO," "Skunk2 Racing," "MBRP performance exhaust." Upon further investigation, we find that they are all related to off-road motocross riding. Some of them are under different category labels on Facebook. For example, "Fox Racing" is labeled as a "retail" company, and it sells motocross and mountain biking gear and apparel. The above results indicate that our approach with engagement data from brands across industries can learn better brand representation and thus a high-quality market structure. To sum up, the power of our deep representation learning for market structure discovery will be at its best when we observe a large scale of user engagement with different brands.

Robustness Check: Visualization Method

Several dimension reduction methods are available for visualizing high-dimensional data. To check whether our derived market structure (300-dimensional representation of brands) is sensitive to the choice of visualization methods, we compare the *t-SNE* with two widely used ones: UMAP (McInnes et al. 2018) and PCA (Wold et al. 1987). UMAP is a recent non-linear visualization technique with its notably fast speed and better preservation of the global structure in high-dimensional data. Principal Component Analysis (PCA) is a canonical linear dimension reduction method. Figure 9 shows the visualization of the derived market structure using UMAP and PCA. Similar to *t-SNE*, the visualization generated by UMAP exhibits clear clustering patterns in that points with same color tend to be in a group (color represents the Facebook category). On the contrary, PCA does not separate industries well. This can be explained by the fact that PCA is a linear transformation method that may not be good at preserving the global and local structure of data in the high-dimensional space. This robustness check confirms that the learned brand representation intrinsically encodes some latent relationships which can be used for discovering market structure.

<Insert Figure 9 Here>

Case Studies on Market Structure Dynamics

Market structure evolves over time and can change dramatically especially under an unexpected industry shock. Whether our proposed method can be adaptively learned is also of interest as it could provide useful insights to marketing practitioners. In this section, we analyze how market structure changes under exogenous shocks by analyzing two case studies: (1) Amazon acquiring Whole Foods, and (2) Tesla introducing the Model 3. We take a before-after strategy where we use data for 3-month pre- and 3-month post the event announcement day and calculate the change in distance from the focal brand (e.g., Amazon and Tesla) to other representative brands that are selected from the same category. The purpose of the event study is to examine how a focal brand relationship with other brands change as a major event occurs. Specifically, for Amazon-Whole Foods, we select several brands from the retail and e-commerce category, and for Tesla, we select several brands from the auto category. This demonstrates that our proposed approach is able to learn effective representation; as a result, the dynamics in market structure are well captured. We calculate the change between focal brand i's representation w_i^b and target brand j's representation w_j^b before and after the specific event using cosine similarity: $cossim\left(w_{i\ after}^{b}, w_{j\ after}^{b}\right) - cossim\left(w_{i\ before}^{b}, w_{j\ before}^{b}\right)$. Therefore, positive numbers indicate similarity increase while negative numbers mean the decrease in similarity.

Amazon acquires Whole Foods

Amazon acquired Whole Foods in June 2017. This acquisition has had significant impact on the grocery and retail industries. It is widely believed that Amazon plans to use its acquisition of Whole Foods to enter into the online grocery delivery business. Amazon and Whole Foods run separate Facebook pages. After the merger of the two firms, we see from Figure 10 that Amazon is more proximal to retail brands as measured by cosine similarity, while the proximity to other relevant brands decreases slightly. For example, the cosine similarity between Amazon and Loews Home Improvement decreases by 0.184. In

contrast, the cosine similarity between Amazon and other super-market brands increases. Among them, proximity of Amazon to Whole Foods increases by 0.202, and increases between Amazon and Kroger by 0.165. As inferred from our data-driven model, Amazon even becomes more proximal to Walmart indicating that Amazon's competitive market structure landscape has shifted. By further examining our data, we find that after the Whole Foods acquisition the number of common users who interact with both Amazon and Whole Foods on their Facebook public pages increases. Some Amazon users posted comments on Whole Foods fan page mentioning Amazon. For example, in a Whole Foods post "Here are 6 New Healthy Products Coming to Whole Foods in March," a user, who had liked an Amazon post earlier, commented "You mean AMAZON… as they bought Whole Foods…right?" This direct link between Amazon and Whole Foods leads the deep Autoencoder to strength the proximity between the two brands. Moreover, in another Whole Foods post, a user who had liked a Kroger post earlier posted "The quality has gone downhill and prices have soared…. You've made Kroger look appealing….'' Although we do not find this user has ever interacted with Amazon before, her interaction with Whole Foods leaves an implicit connection between Amazon and Kroger which could be captured by the deep Autoencoder. In short, after Amazon acquired Whole Foods, online social media users who are Amazon's fans pay more attention to Whole Foods, and users who are fans of other supermarket brands engage more with Whole Foods due to the acquisition event. As a result, the deep Autoencoder captures the dynamics and updates the brand representation accordingly.

<Insert Figure 10 Here>

The acquisition by Amazon has an impact on the market structure of Whole Foods too. In Figure 11, we consider Whole Foods as the focal brand and calculate the change in proximities to other brands before and after the acquisition. Based on the results, we observe that Whole Foods' proximity to other retail brands such as Target, Walmart, and Best Buy increases. Among them, the proximity to Amazon increases the most due to the increased common users between them. In contrast, Whole Foods'

proximity to supermarket brands such as Goya Foods, Enjoy Life Foods, and HelloFresh slightly decreases. Second, the magnitude of change in proximity values is smaller than that of Amazon to other brands. This seems to indicate that the acquisition has less impact on Whole Foods as it is still positioned around other supermarket brands, while Amazon is expanding closer to the grocery retail category.

<Insert Figure 11 Here>

Tesla announces the Model 3

Tesla sells two types of sedans, the Model S and the Model 3. The Model S is a luxury premium sedan with a larger range of acceleration and customization options, while the Model 3 is designed and built as a mass-market affordable electric vehicle. The Model S can cost over $100,000 depending on the configuration, while the Model 3 costs approximately $35,000. After the announcement of the new Model 3, we see that Tesla becomes more distant from luxury car brands and get closer to non-luxury car brands. Examining data from the Auto Gallery, a Southern California premiere luxury and exotic dealership, we can see in Figure 12 that the cosine similarity between Tesla and luxury car brand Maserati decreases by 0.209. Similar trends exist between Tesla and other high-end or luxury car brands such as BMW, Mercedes-Benz, Audi, and so on. Meanwhile, Tesla becomes more proximal to Kia, Mazda, and other more affordable car brands.

<Insert Figure 12 Here>

Testing for Significance

In the above analysis, we compute the distance change between the focal brand (e.g., Amazon or Whole Foods) and other brands, before and after the acquisition. We can see that there is a significant increase in similarity between Amazon and Whole Foods after the acquisition. However, whether this distance change is caused by the acquisition or other unobserved factors, such as the difference of data split and/or noise, still remains unclear. Therefore, we conduct a further analysis by randomly splitting all data before the acquisition into two parts (i.e., d1 and d2, with d1 before d2). We then measure the distance between

Amazon and Whole Foods using d1 and d2 separately. We repeat this process 30 times using different data cuts in the pre-acquisition data. The average distances between the two brands across using all d1s and d2s are 0.228 and 0.232, respectively. The two-tailed t-test on the distance is 0.055, which indicates there is no statistically significant difference between the distances between Amazon and Whole Foods before and after the acquisition in different cuts of the pre-acquisition data. Accordingly, the substantial increase in similarity between Amazon and Whole Foods is not attributed to sample differences.

We perform a similar process on Tesla's introduction of the Model 3. In particular, we choose one non-luxury brand, Mazda, and compute its distances to Tesla before the event using various data splits. The average distances between Mazda and Tesla across using all d1s and d2s are 0.185 and 0.191, respectively, with a *p*-value of 0.076. This seems to indicate there is no statistically significant difference between Mazda and Tesla when the cutting point of data varies before the event. Therefore, we conclude that after Model 3's announcement, Tesla becomes more similar to non-luxury automobile brands on the social media platform. Note that we also conduct analyses on Tesla and other automobile brands and the results are consistent.

External Validation

The brand market structure and case studies described so far help us to reveal brand relationships using online social media users' brand engagement. A question that naturally arises is the extent to which online social media users' brand engagement aligns with other source of data. In this section, we collect search data from Google and derive market structure from the search data, in order to compare the external validity of proposed approach.

External data: We use the Google Search Trend data. It provides an interest score for every search query across regions and languages, as measured by an aggregated search volume over time. A higher interest score means that queries are more popular in a specific region and time. Google search trend data has been widely used by industry (Shimshoni et al. 2015) and academia (Choi and Varian 2012, Du and

Kamakura 2012, Kim and Hanssens 2017, Stephen and Galak 2012) to address marketing and economic problems, e.g. competitive analysis. Researchers also show that this score is consistent with consumers' purchase interest at large (Choi and Varian 2012, Du and Kamakura 2012).

To obtain a relative popularity for every pair of brands, we make a search query consisting of two brand names, for example, "Toyota BMW" or "BMW Toyota" for the brands Toyota and BMW. Note that this simple strategy might miss some search volumes when the query has other words related to brands, such as "Camry," "X5 series." For every brand pair, we can obtain an interest score returned by Google. For example, in the region of United States in 2017, the search interest score is 13 and 85 for the query "Toyota BMW" and "Toyota Honda," respectively. This also indicates that consumers at large are more interested in searching Toyota and Honda together, as compared to searching Toyota and BMW together.

External Validity 1: In the first validity exercise, we focus on the Airline industry and the derived market structure. We have 19 airline brands in our dataset, including United States domestic airlines and International airlines. For every brand pair, we first obtain Google search interest score in the region of United States in 2017 (the same as our engagement data period). Then following previous work (Netzer et al. 2012), we calculate the similarity between two brands A and B as $sim(A,B) = \frac{interest(A,B)}{\sum_{b \in S} interest(b,B)}$, where S is the set of all brands (e.g. 19 here). Netzer et al. use the co-occurrence of two brands in an online discussion forum, instead of a Google search interest score.

Meanwhile, we also calculate similarity for every pair of 19 airline brands using 300-dimensional vectors derived from our deep network representation learning on the engagement data. Cosine similarity is applied.

To check whether two above similarity systems are well aligned, we calculate their Pearson's two-tailed correlation between two sets of 361 (=19*19) similarity scores. It is significantly highly-correlated

($r = 0.630, p = 0.0000$). This indicates that our social engagement based market structure is highly correlated with that derived from Google search trend. Since prior studies have shown that the Google search data trend has a high correlation with consumer's actual purchase, we can conclude that users' social engagement with brands may also contain valuable information for deriving brand relationships.

External Validity 2: We now validate the case study, Amazon acquiring WholeFoods, using Google search trend data. Similar to the first external validity exercise, we choose a total of 29 "retail" brands plus Amazon, (Amazon, Walmart, Target, Macys, Best Buy, Walgreens, Lowes, Whole Foods, IKEA, Sears, 7-Eleven, Dollar General, Sams Club, Dollar Tree, CVS Pharmacy, ALDI, Barnes Noble, Costco, Kroger, Meijer, Safeway, Office Depot, Rite Aid, Albertsons, ShopRite, and The Fresh Market) and obtain their interest scores for every brand pair in the region of United States in 2017. Note that we exclude some small "retail" brands such as Goya Foods, since their Google co-search interest scores with other brands are mostly 0, indicating not enough search data for the brand.

The Pearson's two-tailed correlation between two sets of 900 (=30*30) similarity scores is significantly high, ($r = 0.675, p = 0.0000$) for before acquisition, and ($r = 0.758, p = 0.0000$) for after acquisition. This result confirms the external validity of our social engagement based method. We observe that for Amazon, the most similar brands are Barnes & Noble, Macys and Best Buy before acquisition. After acquisition, the most similar brands are Whole Foods, Barnes & Noble and Macys. For Whole Foods, the most similar brands are The Fresh Market, Albertsons and ShopRite before acquisition. After acquisition, the most similar brands are The Fresh Market, Amazon and Safeway.

We obtain further search interest data for one year after the acquisition (June 2017 to June 2018) because we wanted to examine whether the market structure change is sustained for a long period after the acquisition announcement. For Amazon, the most similar brands are still "Whole Foods," "Barnes & Noble" and "Macys." Other grocery "retail" brands such as Kroger, The Fresh Market become more similar to Amazon than that before acquisition. For Whole Foods, the most similar brands are "The Fresh

Market," "Safeway," "ShopRite" and "Amazon." This indicates that for Amazon, the acquisition impact holds for the extended period of analysis, since Whole Foods is still its most similar brand among these retailer brands. It seems that the acquisition has less impact on Whole Foods as it is still positioned around other supermarket brands. All findings are consistent with our case study using social engagement data providing external validity to our results.

Implications and Conclusion

As our proposed approach handles a larger number of brands and millions of user engagement data across these brands, the results are very useful for brand managers to get a gestalt view of the relationships across thousands of brands. The visualization of potentially overlapping product-market boundaries across many categories helps managers to identify latent threats and potential opportunities which cannot be done with extant methods. For example, for Southwest, is Airfarewatchdog a potential competitor who might draw visitors away from Southwest or is it a complementor who would increase visits to Southwest? Having identified the overlapping market with Airfarewatchdog, Southwest could invest more attention to evaluate the exact nature of this relationship. If Airfarewatchdog is a competitor, then Southwest might focus on developing strategies to differentiate itself and channel visitors to its website exclusively. If it is a complementor, then Southwest might run display ad campaigns on Airfarewatchdog's website. Also, given Hyatt is closely associated with Southwest with common users who "like" both brands, Southwest could run mutually beneficial joint promotions with Hyatt. Identifying such unusual or unexpected insights is the greatest advantage of our approach.

Another important strategic use of our market structure maps is to identify competitors and complementors across industries and track how these relationships change over time. While Hoberg and Phillip (2016) use text analysis to 10-K statements to identify such grouping based on product descriptions that the firms provide, we provide a more dynamic structure based on actual customer/user social media activities. Moreover, our market structure map is more forward-looking and predictive of

emerging competition and complementors and more proactive than those based on 10-K statements, which can be viewed as reactive. Since Hoberg and Phillips (2010) show that merging firms with more similar product descriptions in their 10-Ks results in more successful outcomes, using our market structure maps to identify merger and acquisitions targets (firms sharing common users) may have similar benefits. For example, given that Kawasaki USA, a motorsport vehicle manufacturer, is proximal to Audi USA, is there a benefit for Audi USA to acquire Kawasaki USA?

The power of our method lies in its ability to capture the dynamic changes in market structure. Since the maps are based on the analysis of big data which can be collected in a relatively short window of time, our methodology can be used to track changes in their relative position when firms introduce new products, new promotions, and new marketing initiatives. The case studies that we highlighted provide good illustrations of this. Additionally, although we have not analyzed this in the paper, firms can deploy our method to enhance their social network-based marketing efforts by better targeting specific potential customers, since user nodes in the network are also learned and represented as vectors in the same multi-dimensional space as brands. Our link prediction design demonstrates a possible utilization for targeting. Lastly, our proposed method is generalizable to other similar platforms if we can construct a heterogeneous user-brand network from public fan pages' engagement data.

Our research has some limitations. First, our analysis is conducted on one social network, Facebook. Even though Facebook is one of the largest online social networks that has billions of users and thousands of brands, it is likely that users on different platforms may exhibit different engagement behavior and some of the research findings may not be generalized to other platforms. For example, it is reported that Instagram users and Facebook users have different age groups[6]. We could apply the same technique to other social media platforms and compare findings. When it comes to the dynamic market structure analysis, we generate a series of networks at each give time window. Our current analysis treats these

[6] https://www.statista.com

networks as equally important. In fact, the networks at an earlier stage might affect subsequent networks, because user-brand interactions might be dependent on their prior activities. Incorporating dependency among networks in a temporal way into the representation learning algorithm can improve market structure analysis, which we leave as our future work. Finally, each link in the user-brand network is created when the user engages with the brand on the public page. Facebook has introduced various reaction emotions to the platform to allow users interact with brands in different ways, such as ‘Like,’ ‘Love,’ ‘Haha,’ ‘Wow,’ ‘Sad’ and ‘Angry.’ Future work can build a multi-relation network to deeply capture user-brand engagement heterogeneity.

References

Agrawal A, Gans J, Goldfarb A (2018) *Prediction machines: the simple economics of artificial intelligence* (Harvard Business Press).

Bergen M, Peteraf MA (2002) Competitor identification and competitor analysis: a broad-based managerial approach. *Managerial and Decision Economics* 23(4–5):157–169.

Bergstra J, Bengio Y (2012) Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* 13(1):281–305.

Bhojraj S, Lee CM, Oler DK (2003) What's my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research* 41(5):745–774.

Cao Y, Wang L (2017) Automatic selection of t-SNE Perplexity. *arXiv preprint arXiv:1708.03229*.

Choi H, Varian H (2012) Predicting the present with Google Trends. *Economic record* 88:2–9.

Culotta A, Cutler J (2016) Mining Brand Perceptions from Twitter Social Networks. *Marketing Science* 35(3):343–362.

Day GS, Shocker AD, Srivastava RK (1979) Customer-Oriented Approaches to Identifying Product-Markets. *Journal of Marketing* 43(4):8–19.

Du RY, Kamakura WA (2012) Quantitative Trendspotting. *Journal of Marketing Research* 49(4):514–536.

Erdem T (1996) A Dynamic Analysis of Market Structure Based on Panel Data. *Marketing Science* 15(4):359–378.

Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J (2019) A guide to deep learning in healthcare. *Nature medicine* 25(1):24.

Gabel S, Guhl D, Klapper D (2019) P2V-MAP: Mapping Market Structures for Large Retail Assortments. *Journal of Marketing Research* 56(4):557–580.

Hanssens DM, Pauwels KH, Srinivasan S, Vanhuele M, Yildirim G (2014) Consumer attitude metrics for guiding marketing mix decisions. *Marketing Science* 33(4):534–550.

Hoberg G, Phillips G (2016) Text-based network industries and endogenous product differentiation. *Journal of Political Economy* 124(5):1423–1465.

Ilhan BE, Kübler RV, Pauwels KH (2018) Battle of the Brand Fans: Impact of Brand Attack and Defense on Social Media. *Journal of Interactive Marketing* 43:33–51.

Jacobs G, O'Neill C (2003) On the reliability (or otherwise) of SIC codes. *European Business Review* 15(3):164–169.

John DR, Loken B, Kim K, Monga AB (2006) Brand concept maps: A methodology for identifying brand association networks. *Journal of marketing research* 43(4):549–563.

Kannan PK, Sanchez SM (1994) Competitive Market Structures: A Subset Selection Analysis. *Management Science* 40(11):1484–1499.

Kannan PK, Wright GP, Worobetz ND (1991) Testing for competitive submarkets. *International Journal of Research in Marketing* 8(3):187–203.

Kim H, Hanssens DM (2017) Advertising and word-of-mouth effects on pre-launch consumer interest and initial sales of experience products. *Journal of Interactive Marketing* 37:57–74.

Kim JB, Albuquerque P, Bronnenberg BJ (2011) Mapping online consumer search. *Journal of Marketing Research* 48(1):13–27.

Kübler RV, Colicev A, Pauwels KH (2019) Social Media's Impact on the Consumer Mindset: When to Use Which Sentiment Extraction Tool? *Journal of Interactive Marketing*.

Kuksov D, Shachar R, Wang K (2013) Advertising and consumers' communications. *Marketing Science* 32(2):294–309.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444.

Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. *Journal of Marketing Research* 48(5):881–894.

Levitt T (1960) Marketing myopia. *London: Boston*.

Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. *Journal of the American society for information science and technology* 58(7):1019–1031.
Maaten L van der, Hinton G (2008) Visualizing data using t-SNE. *Journal of Machine Learning Research* 9(Nov):2579–2605.
McInnes L, Healy J, Melville J (2018) UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*.
Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*. 3111–3119.
Moe WW (2006) An empirical two-stage choice model with varying decision rules applied to internet clickstream data. *Journal of Marketing Research* 43(4):680–692.
Mukherjee A, Liu B, Glance N (2012) Spotting fake reviewer groups in consumer reviews. *Proceedings of the 21st international conference on World Wide Web*. (ACM), 191–200.
Nam H, Joshi YV, Kannan P k. (2017) Harvesting Brand Information from Social Tags. *Journal of Marketing* 81(4):88–108.
Naylor RW, Lamberton CP, West PM (2012) Beyond the "like" button: The impact of mere virtual presence on brand evaluations and purchase intentions in social media settings. *Journal of Marketing* 76(6):105–120.
Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine Your Own Business: Market-Structure Surveillance Through Text Mining. *Marketing Science* 31(3):521–543.
Novak TP (1993) Log-linear trees: models of market structure in brand switching data. *Journal of Marketing Research* 30(3):267–287.
Pauwels K, Van Ewijk B (2013) Do online behavior tracking or attitude survey metrics drive brand sales? An integrative model of attitudes and actions on the consumer boulevard. *Marketing Science Institute Working Paper Series* 13(118):1–49.
Pelletier MJ, Horky AB (2015) Exploring the Facebook Like: a product and service perspective. *Journal of Research in Interactive Marketing*.
Pereira HG, de Fátima Salgueiro M, Mateus I (2014) Say yes to Facebook and get your customers involved! Relationships in a world of social networks. *Business Horizons* 57(6):695–702.
Ringel DM, Skiera B (2016) Visualizing Asymmetric Competition Among More Than 1,000 Products Using Big Search Data. *Marketing Science* 35(3):511–534.
Schwarz G (1978) Estimating the dimension of a model. *The annals of statistics* 6(2):461–464.
Shimshoni Y, Efron N, Fink M, Segalis E, Patton B, Levin M, Neufeld M, Bar-Lev N, Matias Y, Tamir N (2015) Campaign and competitive analysis and data visualization based on search interest data.
Shugan SM (1987) Estimating brand positioning maps using supermarket scanning data. *Journal of Marketing Research* 24(1):1–18.
Slotegraaf RJ, Pauwels K (2008) The impact of brand equity and innovation on the long-term effectiveness of promotions. *Journal of Marketing Research* 45(3):293–306.
Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15(1):1929–1958.
Stephen AT, Galak J (2012) The effects of traditional and social earned media on sales: A study of a microlending marketplace. *Journal of marketing research* 49(5):624–639.
Timoshenko A, Hauser JR (2019) Identifying customer needs from user-generated content. *Marketing Science* 38(1):1–20.
Urban GL, Johnson PL, Hauser JR (1984) Testing competitive market structures. *Marketing Science* 3(2):83–112.

Van Vlasselaer V, Eliassi-Rad T, Akoglu L, Snoeck M, Baesens B (2016) Gotcha! Network-based fraud detection for social security fraud. *Management Science* 63(9):3090–3110.

Wold S, Esbensen K, Geladi P (1987) Principal component analysis. *Chemometrics and intelligent laboratory systems* 2(1–3):37–52.

Yang Y, Lichtenwalter RN, Chawla NV (2015) Evaluating link prediction methods. *Knowledge and Information Systems* 45(3):751–782.

Zahedi FM, Abbasi A, Chen Y (2015) Fake-website detection tools: Identifying elements that promote individuals' use and enhance their performance. *Journal of the Association for Information Systems* 16(6):448.

Zaltman G, Coulter RH (1995) Seeing the voice of the customer: Metaphor-based advertising research. *Journal of advertising research* 35(4):35–51.

Zhang K, Bhattacharyya S, Ram S (2016) Large-Scale Network Analysis for Online Social Brand Advertising. *MIS Quarterly* 40(4).

Table 1: Comparison of different types of work on market structure discovery

	Primary/Survey Data	**Text Mining**	**Social Tag-based**	**Search Data**	**Shopping Data**	**Social Engagement**
Data Volume	Small	Large	Large	Large	Very large	Very large
Data Veracity	Authentic	Noisy	Moderately noisy	Moderately noisy	Authentic	Moderately noisy
Privacy Preserving	Yes	Yes	Yes	No (need to insert a tracking pixel)	Yes	Yes
Data Availability	Low (need to do survey)	High (publicly available)	High (publicly available)	Low (need to insert a tracking pixel)	Low (need to partner with retailers)	High (publicly available)
Data pre-processing cost	Low (use consideration set directly)	High (text mining is error-prone)	High (text mining is error-prone)	Low (use consideration set directly)	Low (use product co-occurrence)	Low (use network raw data)

Table 2: Summary of difference among extant literature on market structure discovery

	Kim et al. 2011	**Lee and Bradlow 2011**	**Netzer et al. 2012**	**Ringel and Skiera 2016**	**Culotta and Cutler 2016**	**Nam, Joshi and Kannan 2017**	**Gabel et al. 2019**	**Our study**
Objective	To visualize user search behavior and understand market structure	To visualize competitive market structure using text mining on customer review	To visualize competitive market structure using text mining on forum discussion	To understand asymmetric competition in the product categories	To infer attribute-specific brand ratings	To analyze user generated tags for marketing research	To leverage NLP and ML for analyzing market structure	To propose a novel deep network representation learning framework for market structure
Homogeneous vs. Heterogeneous	Homo	Homo	Homo	Homo	Homo	Homo	Homo	Hetero

Brands/Products	62 products, 4 brands	9 brands	169 products, 30 brands	1,124 products	200 brands	7 brands	133 categories, 30,763 products	5,478 brands
Consumers/Users	N.A.	N.A.	76,587	100,000+	14.6 million	N.A.	N.A.	25,992,832
Data sources	Amazon	Customer review at Epinions	Online discussion forum	Product comparison website	Twitter	Social tagging platform Delicious	Retailer	Facebook public fan page
Data type	Consumer search	Text	Text	Consumer search	Network	Social tags	Shopping baskets	Network
Brand association methodology	Consideratio n set	Text-mining	Text-mining	Consideratio n set	Network learning	Network learning	Network learning	Network learning
Asymmetry	Yes	No	No	Yes	No	No	Yes	Yes
Dynamic	No	No	No	No	No	Yes	No	Yes
Dimension reduction	Yes	Yes	No	No	No	Yes	Yes	Yes
External validation	N.A.	N.A.	Purchase data, survey	Survey	Survey	Brand concept map (survey)	N.A.	Event study, link prediction
Privacy preserve	Yes	Yes	Yes	No (need to insert a tracking pixel)	Yes	Yes	Yes	Yes
Data availability	Low (need to collect data daily)	High (publicly available)	High (publicly available)	Low (need to insert a tracking pixel)	High (publicly available)	High (publicly available)	Low (need to partner with retailers)	High (publicly available)
Data preprocessing cost	Low (use consideration set directly)	High (text mining is error-prone)	High (text mining is error-prone)	Low (use consideration set directly)	Low (use network raw data)	Low (tags are well defined)	Low (use product co-occurrence)	Low (use network raw data)

Table 3: Data description and statistics

Number of brands	5,478
Number of users	25,992,832
Number of unique user-brand interactions	36,927,613
Number of like interactions	87,876,623
Number of unique user-brand like interactions	29,611,805
Number of comment interactions	18,703,549
Number of unique user-brand comment interactions	7,612,358
Total number of user-brand interactions	106,580,172

Table 4: Performance comparison for different models. The number of randomly selected users is *N*=100.

precision@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand network	Shallow model	0.400	0.262	0.132	0.078	0.022	0.012	0.001
		(0.109)	(0.023)	(0.018)	(0.008)	(0.002)	(0.000)	(0.000)
	Deep model	0.410	0.271	0.139	0.082	0.023	0.014	0.001
		(0.092)	(0.027)	(0.020)	(0.009)	(0.003)	(0.001)	(0.000)
Heterogenous brand-user network	Shallow model	0.430	0.291	0.157	0.095	0.028	0.018	0.001
		(0.102)	(0.030)	(0.024)	(0.008)	(0.005)	(0.002)	(0.000)
	Deep model	**0.52*** **	**0.322*** **	**0.173*** **	**0.124*** **	**0.034*** **	**0.028*** **	**0.001*** **
		(0.092)	(0.022)	(0.051)	(0.011)	(0.008)	(0.001)	(0.000)
recall@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand network	Shallow model	0.031	0.260	0.488	0.602	0.828	0.918	0.996
		(0.008)	(0.002)	(0.060)	(0.050)	(0.036)	(0.016)	(0.005)
	Deep model	0.032	0.275	0.505	0.621	0.832	0.912	0.997
		(0.013)	(0.032)	(0.054)	(0.047)	(0.049)	(0.032)	(0.003)
Heterogenous brand-user	Shallow model	0.037	0.287	0.521	0.637	0.870	0.935	0.998
		(0.015)	(0.065)	(0.074)	(0.045)	(0.023)	(0.047)	(0.000)

network		**0.056*** **	**0.311*** **	**0.582*** **	**0.686*** **	**0.897*** **	**0.967*** **	**0.999** **
	Deep model	(0.013)	(0.035)	(0.077)	(0.054)	(0.078)	(0.024)	(0.002)

Table 5: Performance comparison for different models. The number of randomly selected users is *N*=1,000.

precision@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand network	Shallow model	0.460 (0.132)	0.387 (0.112)	0.331 (0.021)	0.291 (0.012)	0.130 (0.004)	0.078 (0.003)	0.012 (0.000)
	Deep model	0.490 (0.020)	0.393 (0.003)	0.332 (0.018)	0.295 (0.017)	0.131 (0.003)	0.078 (0.003)	0.012 (0.000)
Heterogenous brand-user network	Shallow model	0.500 (0.102)	0.422 (0.060)	0.344 (0.022)	0.320 (0.072)	0.162 (0.010)	0.087 (0.017)	0.012 (0.000)
	Deep model	**0.522*** ** (0.092)	**0.436*** ** (0.040)	**0.365*** ** (0.012)	**0.355*** ** (0.035)	**0.187*** ** (0.014)	**0.091*** ** (0.047)	**0.013*** ** (0.000)
recall@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand network	Shallow model	0.031 (0.008)	0.033 (0.021)	0.128 (0.008)	0.223 (0.008)	0.509 (0.013)	0.607 (0.013)	0.915 (0.008)
	Deep model	0.032 (0.005)	0.035 (0.047)	0.131 (0.018)	0.226 (0.011)	0.510 (0.010)	0.605 (0.015)	0.921 (0.007)
Heterogenous brand-user network	Shallow model	0.049 (0.022)	0.056 (0.009)	0.241 (0.012)	0.365 (0.010)	0.549 (0.012)	0.658 (0.024)	0.981 (0.015)
	Deep model	**0.049*** ** (0.009)	**0.076*** ** (0.003)	**0.352*** ** (0.010)	**0.412*** ** (0.007)	**0.584*** ** (0.009)	**0.743**** (0.008)	**0.990*** ** (0.002)

Table 6: Performance comparison for different sizes of training set. The number of randomly selected users is N=1,000.

precision@1000		10%	30%	50%	70%	90%	100%
Homogeneous brand-brand network	Shallow model	0.103	0.195	0.248	0.263	0.282	0.291
		(0.012)	(0.008)	(0.008)	(0.012)	(0.015)	(0.012)
	Deep model	0.097	0.190	0.248	0.267	0.284	0.295
		(0.042)	(0.010)	(0.021)	(0.031)	(0.023)	(0.017)
Heterogenous brand-user network	Shallow model	0.143	0.225	0.256	0.283	0.312	0.320
		(0.015)	(0.031)	(0.042)	(0.008)	(0.052)	(0.072)
	Deep model	**0.183*** **	**0.242*** **	**0.273*** **	**0.301*** **	**0.337*** **	**0.355*** **
		(0.024)	(0.032)	(0.037)	(0.012)	(0.032)	(0.035)
recall@1000		10%	30%	50%	70%	90%	100%
Homogeneous brand-brand network	Shallow model	0.080	0.153	0.193	0.203	0.219	0.223
		(0.009)	(0.006)	(0.006)	(0.007)	(0.011)	(0.008)
	Deep model	0.075	0.150	0.194	0.204	0.220	0.226
		(0.005)	(0.010)	(0.007)	(0.003)	(0.005)	(0.011)
Heterogenous brand-user network	Shallow model	0.108	0.179	0.223	0.257	0.271	0.281
		(0.031)	(0.018)	(0.013)	(0.026)	(0.017)	(0.010)
	Deep model	**0.124*** **	**0.198*** **	**0.24*** **	**0.289*** **	**0.314*** **	**0.352*** **
		(0.009)	(0.008)	(0.019)	(0.029)	(0.008)	(0.007)

Table 7: Top 10 proximal brands to each focal brand

Focal brand		United	Southwest Airlines	Audi USA	Nissan
Rank	1	American	JetBlue	Mercedes-Benz USA	Mazda
	2	Delta	Frontier	BMW USA	Toyota
	3	Lufthansa	Allegiant	Land Rover	Volkswagen
	4	Southwest	Delta	Lexus	Kia Motors America
	5	Alaska	Alaska	Chevrolet Camaro	Subaru of America
	6	All Nippon	United	Maserati USA	Chrysler
	7	Air China	Airfarewatchdog	Kawasaki USA	FIAT
	8	LATAM	American	Firestone Tires	Jaguar
	9	Air New Zealand	Virgin America	Tesla	Alfa Romeo
	10	Airfarewatchdog	Hyatt	Ram Trucks	KLIM

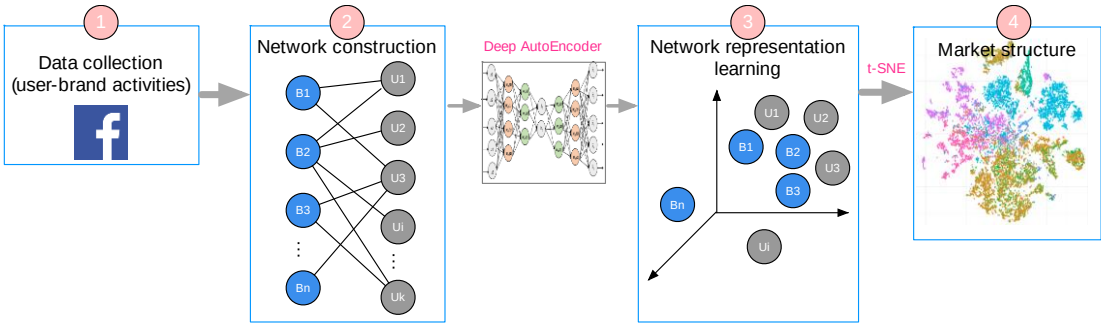


Figure 1: The overall framework of the proposed deep network representation learning

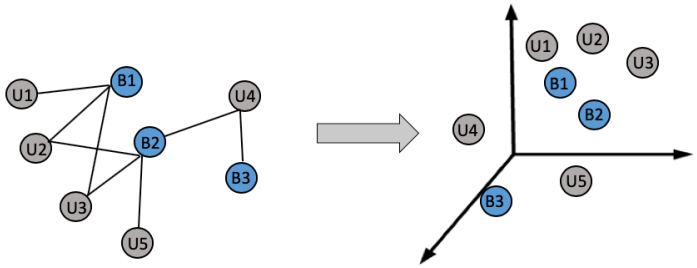


Figure 2: An illustration of deep network representation learning

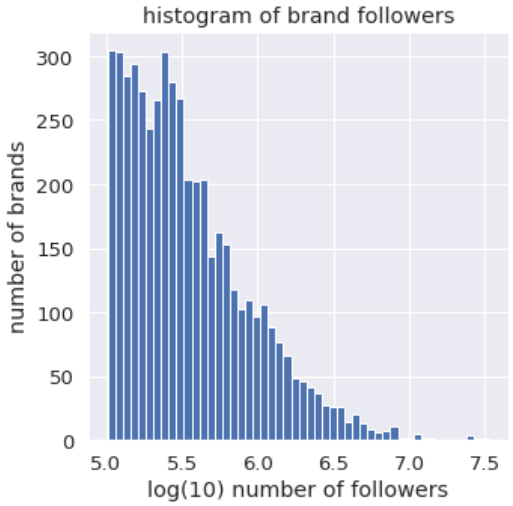


Figure 3: Histogram of number of followers of 5,478 Facebook brands

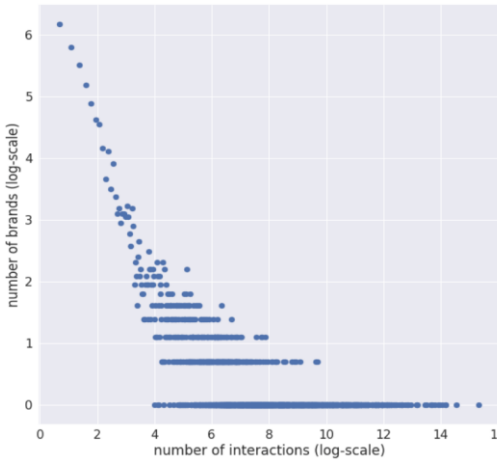


Figure4: Degree distribution of brands in the user-brand network

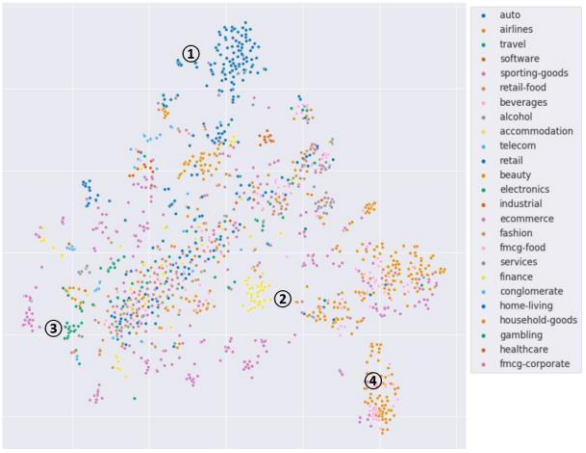


Figure 5: The global structure among brands in our Facebook data

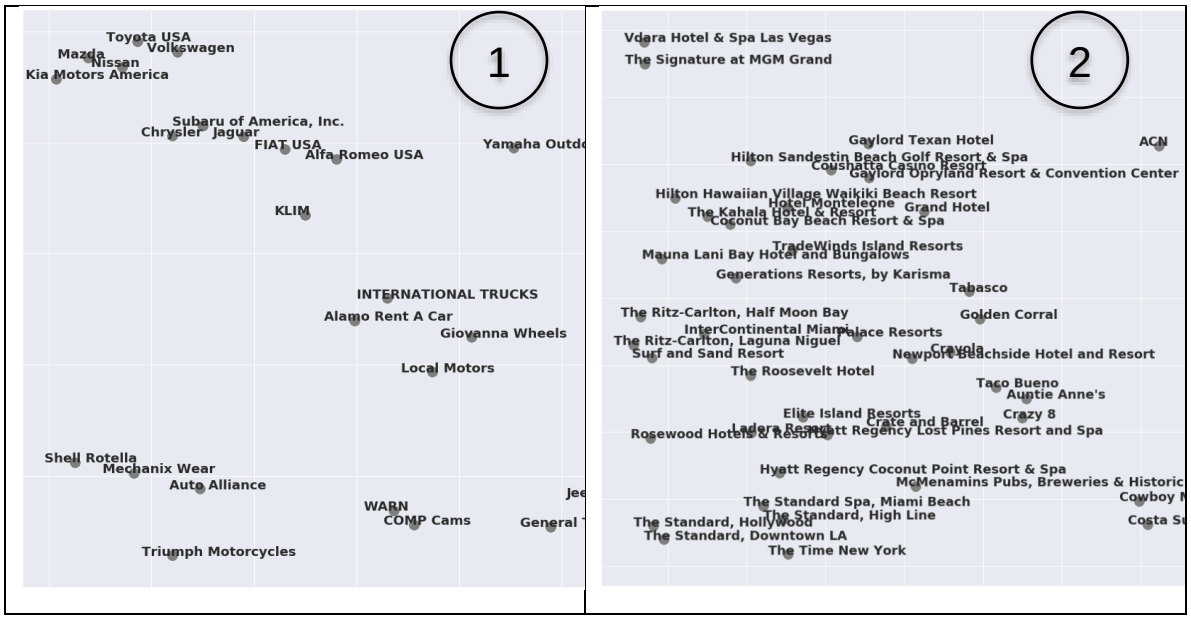


Figure 6: Zooming-in on Clusters 1, 2, 3, and 4

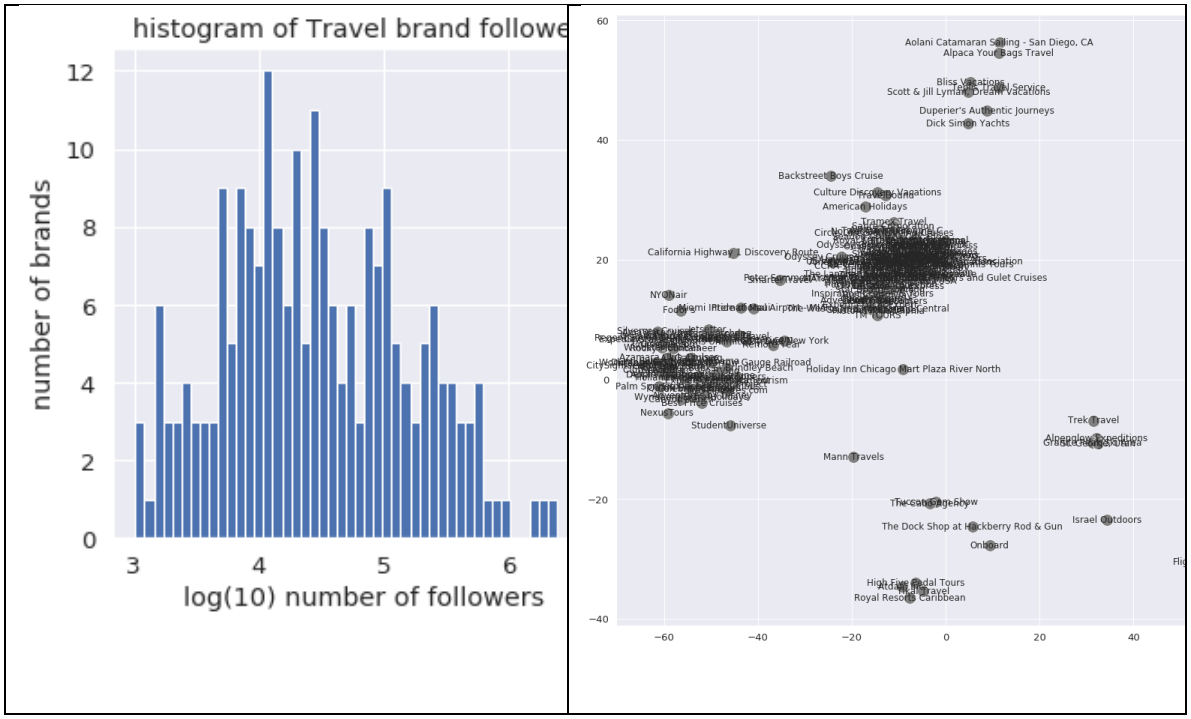


Figure 7: Distribution of brand size (log10 base) and the visualization of market structure of 242 small brands

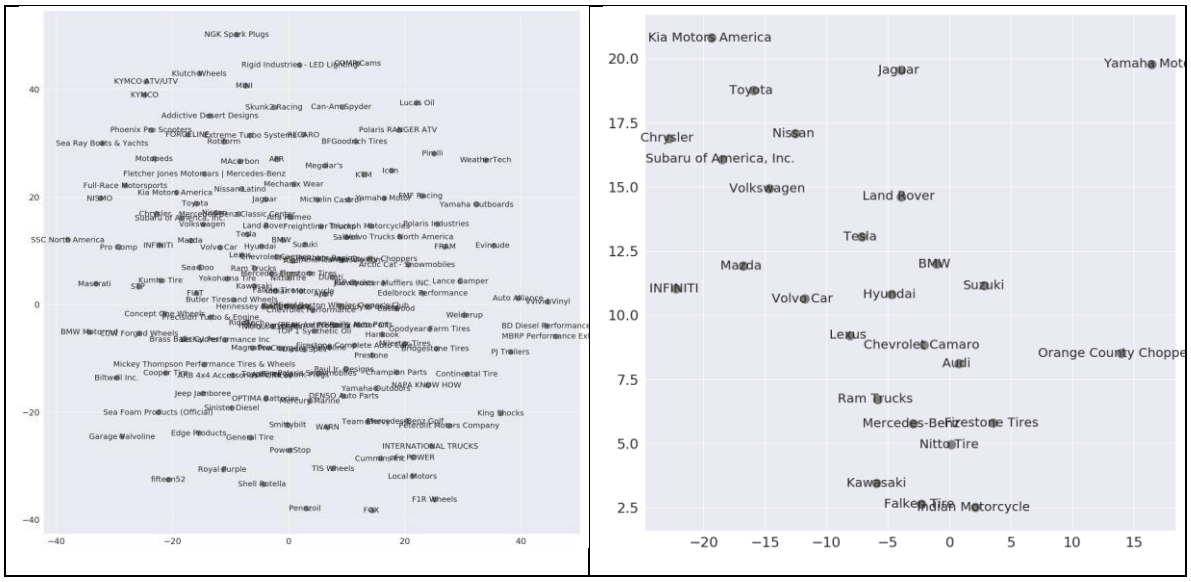


Figure 8: Visualization of market structure of using engagement data only from "auto" brands: all 163 auto brands (left) and 27 large auto brands with more than 1 million followers (right)

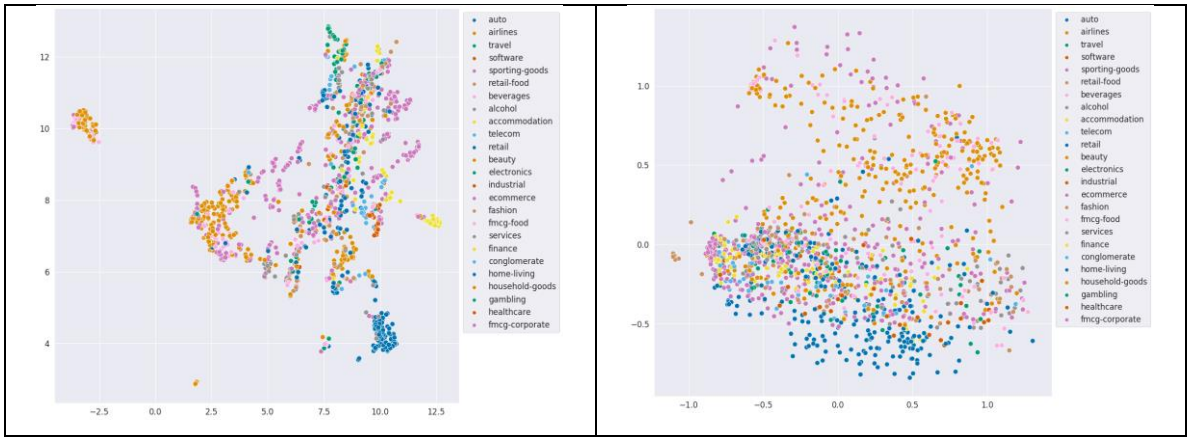


Figure 9: Visualization of market structure using UMAP (left) and PCA (right)

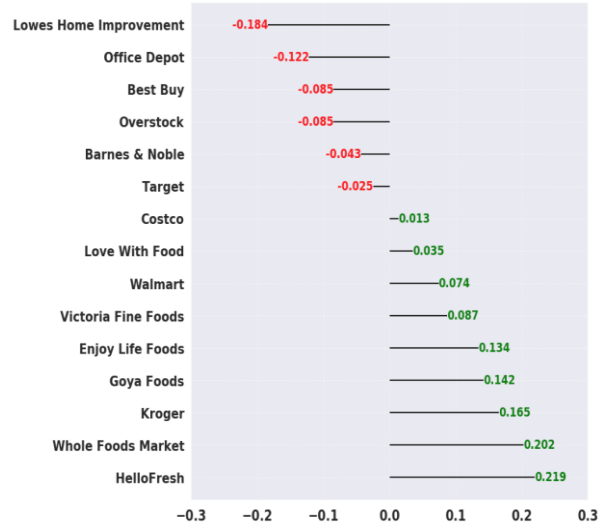


Figure 10: Similarity change of Amazon to other brands in retail and e-commerce industry

Figure 11: Similarity change of Whole Foods to other brands in retail and e-commerce industry

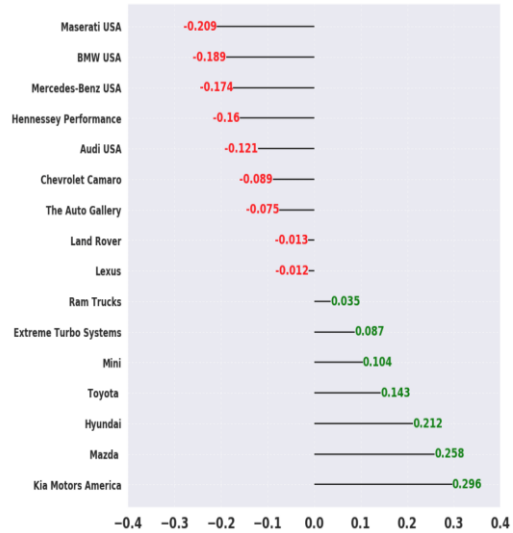


Figure 12: Similarity change of Tesla to other selected brands in the auto industry

Appendixes: Identifying Market Structure: A Deep Network Representation Learning of Social Engagement

A1: Deep Autoencoder

In this study, the network representation, also known as network embedding, is learned through a deep Autoencoder, an unsupervised learning model consisting of two joint components, an encoder, and a decoder. The encoder, implemented by a deep fully-connected feed forward neural network, is a compressor that transforms the input data into a latent representation (e.g., a low-dimensional vector), while the decoder is a reverter that reconstructs the latent representation back to the original input data. As the input data is often high dimensional (in a magnitude of millions), learning effective low dimensional representation (several hundred) in an efficient way while preserving information in the input data as much as possible is not trivial. In our case we study a large user-brand network where each brand (or user) node is originally represented as a one-hot encoding vector that is fed into the Autoencoder for learning a latent low-dimensional vector. To illustrate the Autoencoder in details, we first formally define user-brand heterogeneous network and network representation learning.

Definition 1: user-brand Heterogeneous Network A user-brand heterogeneous network is denoted as $G = (V^b, V^u, E)$, where $V^b = (v_1^b, v_2^b, \dots, v_n^b)$ represents n brand nodes, $V^u = (v_1^u, v_2^u, \dots, v_m^u)$ represents m user nodes, and $E = \{e_{i,j}\}, i \leq m, j \leq n$ represents all links between users and brands. $e_{i,j}$ is the link weight that indicates the frequency of engagement between user i and brand j. Engagement is defined as liking or commenting by a user on a brand's Facebook fan page.

Definition 2: Network Representation Learning Given a user-brand heterogeneous network G, network representation learning aims to learn a mapping function $f: v_i^b, v_j^u \mapsto w_i^b, w_j^u \in R^d$, where $d \ll \min(m, n)$. w_i^b, w_j^u are called brand embeddings and user embeddings, respectively.

The objective of the mapping function is to learn good embeddings so that the brand proximities, brand-user proximities, and user proximities are preserved at maximum. More specifically, given network-like inputs, we tend to preserve the following two network structures into the learned representations.

1. Similarity to neighbors. Our user-brand network is a bipartite network where brand nodes and user nodes are neighbors. A user node and a brand node are connected with a large weight, indicating a strong relationship between them. The similarity from this one-hop connection for all links between users and brands are measured by first-order loss function, denoted as L_{1st}. L_{1st} with weights incorporated incurs a penalty if neighboring nodes are projected far apart, similarly to the idea of Laplacian Eigenmaps (Belkin and Niyogi 2003). Therefore, minimizing L_{1st} is an attempt to preserve local distances; if v_i^b and v_j^u are similar, then w_i^b and w_j^u are close in the embedding space.

$$L_{1st} = \sum_{j=1}^{n}\sum_{i=1}^{m} e_{i,j}\left(w_i^b - w_j^u\right)^2$$

2. Similarity to neighbors of neighbors. In our user-brand network, neighbors' neighbors of a brand node are other brand nodes. Neighbors' neighbors of a user node are other user nodes. If two brands share many common users, their similarity should be high. Similarly, if two users are fans for many common brands, their similarity should be high too. The objective of network representation learning is designed in such a way that a network structure similarity should be well captured. Therefore, to minimize the reconstruction error (denoted as L_{2st}) by compressing the latent information in hidden layers, the Autoencoder has the following objective function measured by second- order loss function, denoted as L_{2nd}.

$$L_{2nd} = \sum_{i=1}^{m}(x_i^{b'} - x_i^b)^2 + \sum_{j=1}^{n}(x_j^{u'} - x_j^u)^2$$

where x_i^b and x_j^u are the input of brand v_i^b and user v_j^u for the deep Autoencoder, respectively. They are represented as an adjacent one-hot encoding vector by all other nodes. The dimensionality of x_i^b and x_j^u equals the total number of brands and users $(m+n)$ in the network. Each element in the vector corresponds to a node in the network. If the node at a particular index connects the brand node v_i^b (or user node v_j^u), the corresponding element is marked as the engagement frequency, and as 0 otherwise. This adjacent representation is a very common way for representing nodes in a network (Liben-Nowell and Kleinberg 2007).

Therefore, our overall objective function is to minimize the sum of first-order loss and second-order loss, as below.

$$L = L_{1st} + L_{2nd} = \sum_{j=1}^{n}\sum_{i=1}^{m} e_{i,j}\left(w_i^b - w_j^u\right)^2 + \lambda\left(\sum_{i=1}^{m}\left(x_i^{b\prime} - x_i^b\right)^2 + \sum_{j=1}^{n}\left(x_j^{u\prime} - x_j^u\right)^2\right)$$

$x_i^{b\prime}$ and $x_j^{u\prime}$ are the output of the deep Autoencoder, which are the reconstructed representation of the input x_i^b and x_j^u, respectively. The hyper-parameter λ plays a role to balance the first-order loss and second-order loss, and its value are tuned using grid search via the link prediction experiment. The essence of a deep Autoencoder is to minimize the reconstruction error between the input and output via deep neural networks. In particular, given input x_i^b, parameters of the intermediate representation for each encoder layer are as follows:

$$w_i^1 = \sigma(W^1 x_i + b^1)$$

$$w_i^k = \sigma\left(W^k w_i^{k-1} + b^k\right), k = 2, \ldots, K$$

After we obtain the intermediate representation w_i^K, the output x_i' can be generated via a reversing operation of the encoder. That is, the network parameters of the k-th layer are shared between the encoder and decoder. The reconstruction process for the decoder layers is as follows:

$$w_i^{K\prime} = \sigma(W^K x_i + b^K)$$

$$x_i' = \sigma(W^1 w^{1\prime} + b^{1\prime})$$

We implement the above deep learning model using the Tensorflow library on Nvidia P100 GPU. Gradient descent is used in optimization and parameter estimation. We also adopt dropout training (Srivastava et al. 2014), a common practice in neural network, to avoid overfitting. In our experiments, we use sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ as the activation function to capture the non-linearity.

Parameter tuning: It is well known that hyperparameter selection plays a critical role in deep learning model performance. In the proposed deep autoencoder network, the key parameters include: the brand representation dimensionality, the number of hidden layers, the number of neurons in each hidden layer, and the parameter λ in the loss function. Another important aspect to consider when selecting hyperparameters is the computational cost. An exhaustive parameter search is infeasible since it would be extremely costly both in computing power and time. For example, training for one parameter setting under using an Nvidia P100 GPU with 16GB GPU memory costs approximately 10 minutes. Therefore, we take a random search strategy (Bergstra and Bengio 2012), which is shown to be more effective and efficient than traditional grid-search for hyperparameter tuning. Specifically, we treat brand representation dimensionality as a uniform random variable in the range of 100 and 500. For the number of hidden layers, we need to make a tradeoff and set to 3, as an autoencoder with more hidden layers has more parameters to train and is prone to overfit, while an autoencoder with fewer hidden layers has less expressive power to learn complex patterns from large data. For the number of neurons in each layer, we do the following. We treat the number of neurons as a uniform random variable in the range of 2000 and 5000, 500 and 1000 for the first and the second hidden layer, respectively. The number of neurons in the third hidden layer is the same as the network representation dimensionality. Finally,

for parameter λ that controls the first-order loss and the second-order loss, since we expect to achieve a balance between network local structure (first-order) and network global structure (second-order), we choose $\lambda = 1$ without further tuning. We use the link prediction experiment to tune the hyperparameters, that is, the hyperparameter combination that achieves the best performance on the validation set is considered optimal and used for subsequent analyses. We split the entire dataset into 80% training, 10% validation and 10% testing. Using this random search, we find that link prediction performance becomes stable when the representation dimensionality is between 250 and 400, the number of neurons in the first and the second hidden layer is 5000 and 1000, respectively. Therefore, the final setting of our deep autoencoder is as follows. The representation dimensionality d is 300. It has three hidden layers in the encoder, i.e., K=3. Each hidden layer has 5,000, 1,000, and 300 neurons, respectively. The decoder uses exactly the same number of hidden layers and neurons, i.e., 300, 1,000, and 5,000.

A2: Link prediction

The link prediction process follows the seminal work (Liben-Nowell and Kleinberg 2007). Let $G_{0,2} = (V_{0,2}^b, V_{0,2}^u, E_{0,2})$ denote a network snapshot during a time period (t_0, t_2). The network $G_{0,2}$ can be chronologically split into two non-overlapping sub-networks $G_{0,1} = (V_{0,1}^b, V_{0,1}^u, E_{0,1})$ and $G_{1,2} = (V_{1,2}^b, V_{1,2}^u, E_{1,2})$. Conventionally, we call $G_{0,1}$ and $G_{1,2}$ training network and testing network, respectively. The overall evaluation process is as follows. First, we train on $G_{0,1}$ to obtain brand representation (and user representation). Second, we randomly select *N* users in the period of (t_0, t_1). For each user, we calculate its proximity to all non-connected brands. We sort all proximity scores for all *N* users and choose top *k* pairs (denoted as *L*) as predicted links. Finally, we evaluate the performance using two standard metrics: $precision@k$ and $recall@k$, defined below. Precision indicates the accuracy of the link prediction algorithm while recall is referred to as the true positive rate or sensitivity. The larger value for both metrics indicates the better performance. Note that we use precision-recall instead of ROC curves (and AUC) because the latter one is not a meaningful metric in the link prediction problem (Yang et al. 2015). In a social network, the ratio between formed links and all possible links is extremely low. For example, on Facebook, a user only interacts with a small number of brands. ROC curve, and its area (AUC) is equivalent to the probability of a randomly selected positive instance appearing above a randomly selected negative instance in the prediction score space. If we treat formed links as positive instance and non-formed links as negative links, due the high sparsity, such probability would be close to 1.0 (the perfect model) even for a mediocre learning model, which can be deceptive. Therefore, we follow prior literature in social network link prediction (Liben-Nowell and Kleinberg 2007) and use precision-recall as evaluation metrics.

$$precision@k = \frac{|L \cap E_{1,2}|}{k}, \quad recall@k = \frac{|L \cap E_{1,2}|}{|E_{1,2}^T|},$$

where $E_{1,2}$ is the set of all newly formed links in $G_{1,2}$. $precision@1$ checks whether a non-connected brand-user pair with the highest proximity in the training period forms a link in the testing network. Note that this evaluation process might be slightly different when it comes to a brand-brand homogenous network where we only have vector representation for brands. To obtain the proximity scores to all non-connected brands for N randomly selected users, we employ a weighted average strategy, a similar idea used in the item-based collaborative filtering framework. For each user u_i, we have all brands that u_i has connected (i.e., b_1, b_2, …, b_m that u_i connects to in $G_{0,1}$). The similarity score between u_i and each non-connected brand b_j is S_{ij}: $S_{ij} = \frac{\sum_{k=1}^{m} S_{kj}}{m}$, where S_{kj} is the similarity between brand b_j and brand b_k that u_i connects to in $G_{0,1}$, m is the number of brands u_i connects to in $G_{0,1}$.

The advantage of analyzing a homogeneous network is an increase in computational efficiency because the network size is dramatically reduced. However, such a simplified operation that converts an original heterogeneous user-brand network into an implicit homogeneous network usually results in decreased performance because some important information encoded in user-brand interactions is completely ignored. In contrast, our deep learning-based approach performs well because it jointly learns optimal representation for both brands and users while preserving latent relationships among brands and users. For the shallow model, we use the singular-value decomposition (SVD) method, which is the essential method in PCA to learn low-dimensional factors of the input data. Given a brand-user engagement matrix, we use SVD to decompose the user-brand interaction matrix M into a lower rank approximation: $M = U\Sigma V^T$, where U conceptually represents how much each user "likes" an underlying dimension, V^T conceptually represents how relevant each underlying dimension is to each brand, and Σ is a diagonal matrix of singular values, which are essentially weights. For the purpose of prediction, we first approximate the original matrix through U, Σ, and V^T and then predict a link to a brand with the highest predicted preferences that the user has not connected.

Algorithm: LINK PREDICTION ALGORITHM

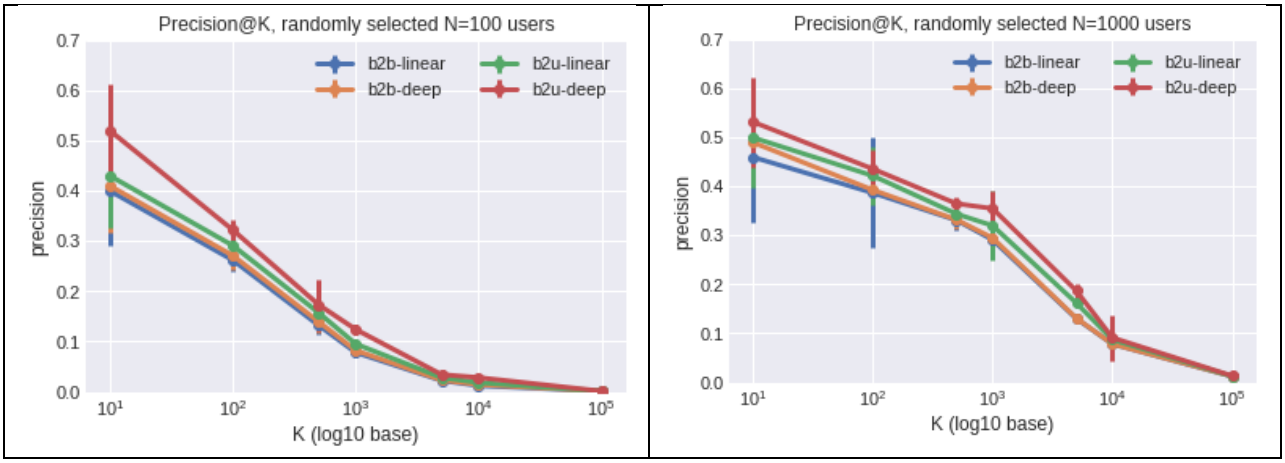
Input: user-brand networks in training and testing; number of randomly selected users: N
k: *precision@k, recall@k*
Output: *precision@k* and *recall@k*

1. Obtain node representation V_i via deep autoencoder for i in 1, …, m (m is the total number of users in training)
2. Select N users at random $U = \{u_1, u_2, \ldots, u_N\}$
3. $S \longleftarrow \Phi$ (initialization)
4. **foreach** user $u_i \in U$ **do**
   foreach brand b_j in training **do**
   $p_{ij} \longleftarrow$ proximity score between u_i and b_j
   S += ($u_i \longleftrightarrow b_j$, p_{ij})
   end
   end
5. $L = \{l_1, \ldots, l_k \mid l_i$ is a user-brand pair$\}$ (top k predicted links based on their proximity scores)
6. $precision@k = \frac{|L \cap E_{1,2}|}{k}$, $recall@k = \frac{|L \cap E_{1,2}|}{|E_{1,2}^T|}$ (see the definition of $E_{1,2}$ and $E_{1,2}^T$ in the Evaluation and Results Section)

Web Appendix: Identifying Market Structure: A Deep Network Representation Learning of Social Engagement

WA1: Link Prediction Precision-recall Performance

In this study, we aggregate user comments and likes as engagements and build a user-brand network. To evaluate the accuracy of learned brand and user network representations, we introduce a novel *link prediction research design*, where we predict the most likely formed links of user-brand engagement in an out-of-sample network given the brand representations and user representations learned from a training network. The experiment results are summarized in Table 4-6 in the main body of the paper. To facilitate a visual presentation of the results, we present the link prediction results in Figure WA1 and Figure WA2. We also demonstrate the impact of different scales of data size on the link prediction performance. Note that in our experiment, we compile a 2x2 research design with two different network structures (homogeneous vs. heterogeneous) and two different algorithms (linear model vs. deep model). We denote *b2b* as the homogeneous brand-brand network, and *b2u* as the heterogeneous brand-user network. Accordingly, four combinations are *b2b-linear*, *b2b-deep*, *b2u-linear*, and *b2u-deep* represent the method that uses linear network learning on the *b2b* network, deep network learning on the *b2b* network, linear network learning on the b2u network, and deep network learning on the *b2u* network, respectively.



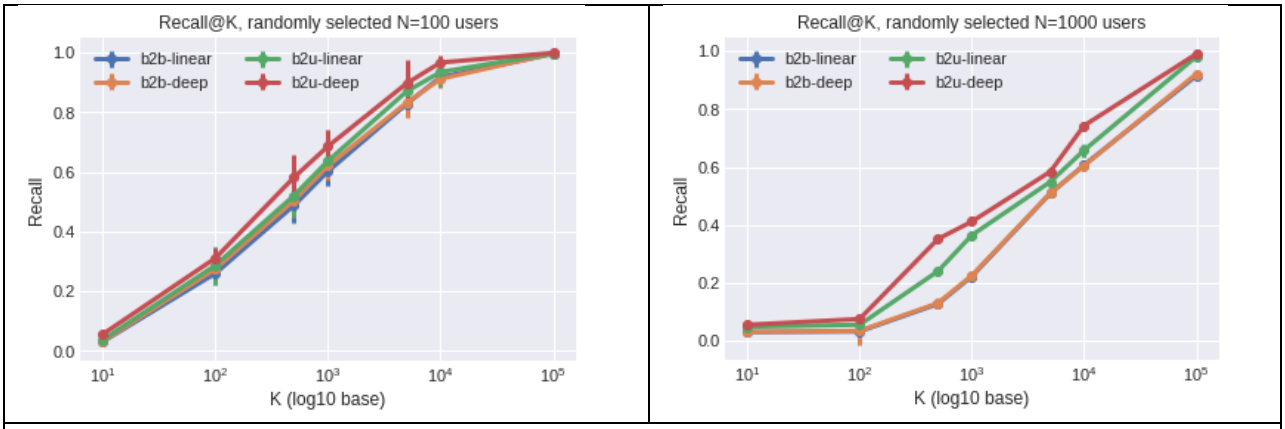


Figure WA1: Link prediction performance, evaluated by *Precision@K*, and *Recall@K*. The X-axis denotes the value of *K*, indicating the top *K* selected user-brand links. The Y-axis denotes the precision and recall number

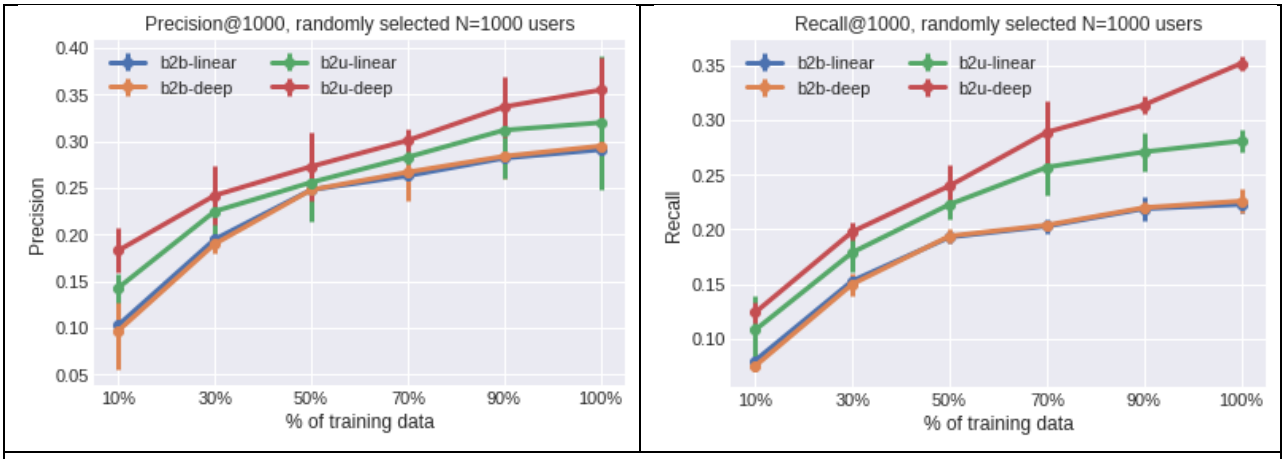


Figure WA2: Link prediction performance, with different scales of training data. The X-axis denotes the percentage of training data used in building and training the user-brand network

WA2: Comment Network and Like Network

On Facebook, users engage with brands in multiple ways, such as liking or commenting. In this study, we aggregate user comments and likes as engagements and build a user-brand network. However, it is also interesting to know what would happen if we construct a user-brand network with only comments or likes. Prior research shows that Facebook likes affect offline customer behavior (Mochon et al. 2017). To have a deeper understanding of learned network representation, we conduct two complimentary link-prediction experiments based on the comment network and the like network. The comment network is constructed between a user and a brand if the user leaves comments on the brand's public page. Similarly, the like network is constructed between a user and a brand if the user likes posts on the brand's public page.

Similar to our previous experiments, we measure the link prediction performance on two metrics *precision@k* and *recall@k.*

We can observe several findings from the results shown in Table WA1 and Table WA2. First, the network representations learned from the like network or the comment network have less predictive power than those learned from the network constructed using both likes and comments. For example, the *precision*@1000 of the comment network, the like network and the like+comment network is 0.168, 0.314 and 0.355, respectively. Similarly, results for the *recall* metric show that the deep network learning is better at capturing the hidden relationships among brands and users with more volume and variety of data. Second, we see that deep network learning approach consistently performs better than linear models in the heterogeneous user-brand network setting, where the performance gain is limited in the homogeneous brand-brand network setting. This indicates that a common practice of reducing heterogeneous networks to homogeneous networks loses important information for learning good representation. Third, we can see that link prediction performance is better for the like network than the comment network. The reasons are two-fold: (1) the like network has more data than the comment network, which facilitates better network representation learning, and (2) the like engagement is more

meaningful than the comment engagement in the market structure discovery. A user liking a brand signals a preference for the brand, while a user commenting on a brand can be a complex signal as the comment may be positive or negative.

Table WA1: Performance comparison for different models on the Like network. The number of randomly selected users is *N*=1,000

precision@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand network	Linear model	0.320 (0.094)	0.279 (0.056)	0.258 (0.008)	0.233 (0.008)	0.127 (0.004)	0.067 (0.001)	0.011 (0.001)
	Deep model	0.323 (0.147)	0.284 (0.082)	0.258 (0.017)	0.235 (0.009)	0.135 (0.014)	0.069 (0.034)	0.011 (0.002)
Heterogeneous brand-user network	Linear model	0.424 (0.035)	0.365 (0.042)	0.312 (0.039)	0.287 (0.008)	0.152 (0.032)	0.087 (0.003)	0.011 (0.000)
	Deep model	**0.486*** ** (0.026)	**0.398*** ** (0.032)	**0.354*** ** (0.023)	**0.314*** ** (0.009)	**0.178*** ** (0.037)	**0.091*** ** (0.004)	**0.011** (0.001)
recall@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand network	Linear model	0.002 (0.001)	0.024 (0.005)	0.111 (0.003)	0.201 (0.006)	0.458 (0.015)	0.563 (0.010)	0.896 (0.006)
	Deep model	0.002 (0.002)	0.025 (0.002)	0.124 (0.011)	0.204 (0.018)	0.476 (0.052)	0.560 (0.023)	0.882 (0.034)
Heterogeneous brand-user network	Linear model	0.041 (0.003)	0.056 (0.004)	0.332 (0.029)	0.350 (0.029)	0.521 (0.075)	0.635 (0.079)	0.911 (0.009)
	Deep model	**0.049*** ** (0.005)	**0.068*** ** (0.006)	**0.350*** ** (0.021)	**0.404*** ** (0.043)	**0.562*** ** (0.037)	**0.663*** ** (0.063)	**0.929*** ** (0.028)

Table WA2: Performance comparison for different models on the Comment network. The number of randomly selected users is *N*=1,000

precision@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand network	Linear model	0.189	0.179	0.156	0.134	0.067	0.045	0.010
		(0.169)	(0.041)	(0.014)	(0.008)	(0.005)	(0.003)	(0.000)
	Deep model	0.189	0.168	0.162	0.137	0.062	0.044	0.010
		(0.097)	(0.019)	(0.052)	(0.010)	(0.032)	(0.002)	(0.001)
Heterogeneous brand-user network	Linear model	0.213	0.192	0.167	0.154	0.122	0.080	0.010
		(0.025)	(0.087)	(0.029)	(0.024)	(0.052)	(0.020)	(0.001)
	Deep model	**0.234*** **	**0.210*** **	**0.173*** **	**0.168*** **	**0.126*** **	**0.088*** **	**0.011***
		(0.045)	(0.023)	(0.067)	(0.019)	(0.033)	(0.002)	(0.002)
recall@k		k=10	k=100	k=500	k=1,000	k=5,000	k=10,000	k=100,000
Homogeneous brand-brand network	Linear model	0.002	0.017	0.068	0.117	0.291	0.393	0.834
		(0.002)	(0.003)	(0.006)	(0.008)	(0.017)	(0.018)	(0.008)
	Deep model	0.002	0.019	0.068	0.114	0.295	0.393	0.842
		(0.001)	(0.012)	(0.022)	(0.032)	(0.042)	(0.053)	(0.012)
Heterogeneous brand-user network	Linear model	0.019	0.042	0.077	0.162	0.333	0.442	0.885
		(0.003)	(0.019)	(0.045)	(0.029)	(0.029)	(0.056)	(0.034)
	Deep model	**0.018**	**0.044** **	**0.082*** **	**0.182*** **	**0.352*** **	**0.453*** **	**0.894*** **
		(0.004)	(0.012)	(0.051)	(0.037)	(0.026)	(0.033)	(0.046)

WA3: Category-level Visualization

Additionally, since each brand in our Facebook data is associated with a category (provided by the brand), we can visualize how the 25 Facebook categories are related. We take a weighted average on learned vectors of all brands within each category to obtain a category vector. It can be considered as a "centroid" of all brands that belong to that category. More formally, given a category C that includes a set of brands $\{b_1, b_2, \dots, b_k\}$, where $\boldsymbol{v_i}$ is the vector representation of each brand b_i, f_i is the number of users who have engaged with b_i. Then, the category representation of C is represented as: $\boldsymbol{v_C} = \sum_{i=1}^{k} \log(f_i)\, \boldsymbol{v}_i$. Once we obtain vector representations for all 25 categories, we visualize them on a two-dimensional space using *t-SNE*, as shown in Figure WA3. We can see that "travel" is next to "airlines", and not surprisingly, "alcohol" is close to "sporting-goods" and "gambling".

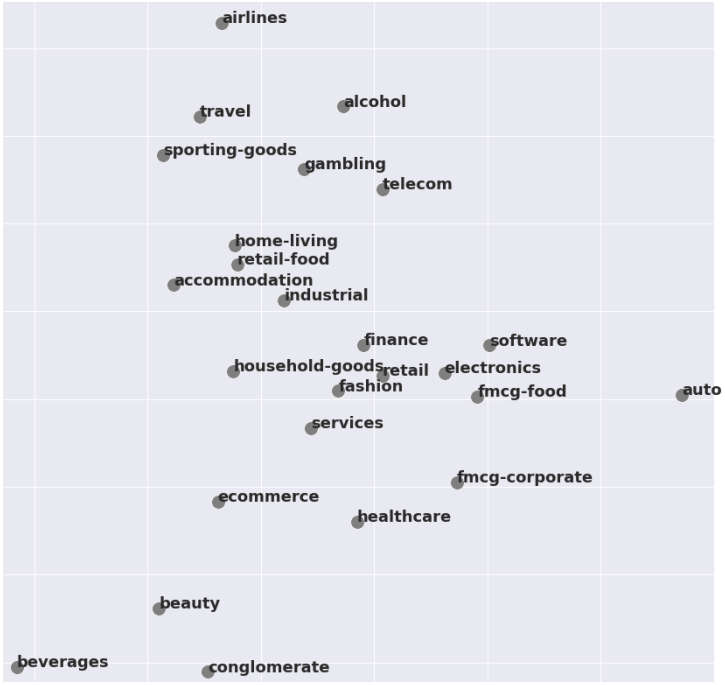


Figure WA3: Visualization of Facebook category structure.